MICROCOPY RESOLUTION TEST CHART
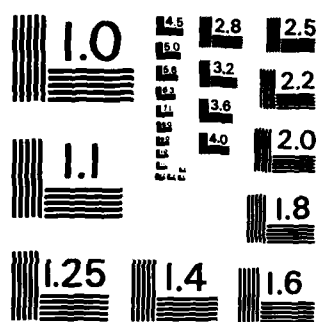
NATIONAL BUREAU OF STANDARDS-1963-A

AD-A147 765

SELECTION OF AIR TRAFFIC CONTROLLERS

Edited by:

S. B. Sells
Texas Christian University

John T. Dailey and Evan W. Pickrel
Federal Aviation Administration

August 1984

DTIC
ELECTE
NOV 1 4 1984
B

84 10 25 073

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FAA-AM-84-2	AD-A147765	

4. Title and Subtitle	5. Report Date *AUG.* ~~June~~ 1984
Selection of Air Traffic Controllers	6. Performing Organization Code
	8. Performing Organization Report No.

7. Author's)
S.B. Sells, J.T. Dailey, & E.W. Pickrel

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Office of Aviation Medicine Federal Aviation Administration 800 Independence Avenue, S.W. Washington, D.C. 20591	11. Contract or Grant No.
	13. Type of Report and Period Covered

12. Sponsoring Agency Name and Address	
Federal Aviation Administration Office of Aviation Medicine 800 Independence Avenue, S.W. Washington, D.C. 20591	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

An encyclopedic report on air traffic controller selection research. Eighteen contributors have prepared twenty-five chapters encompassing research over the past 40 years.

A historical review of controller selection research includes an international overview, U.S. research from 1941 to 1963, contributions of the Civil Aeromedical Institute and the Office of Aviation Medicine, and adjustments following the PATCO strike.

A section on job analysis and characteristics of air traffic controllers is followed by six chapters on measurement of air traffic controller performance. These include Terminal, Enroute, and Flight Service Station training program assessment, controller skills tests, dynamic paper-and-pencil simulations for proficiency measurement, and criterion measurement in selection research.

Research leading to the FAA's 1981 ATC selection tests includes chapters on development of the new Multiplex Controller Aptitude Test and Occupational Knowledge Test, personality assessment of ATC applicants, studies from 1972 through 1978 to validate the new selection tests, conformity of the new experimental battery to the Uniform Guidelines on Employee Selection Requirements, and recommendations for adoption of the new battery and further research.

An overview of projected developments in ATC systems technology from now to the year 2000 is used to project changes that will occur in the air traffic controller's future role and function.

17. Key Words	18. Distribution Statement
Air Traffic Controller Selection Aptitude Testing Personnel Research Performance Measurement Personality Assessment	Document is available to the public through the National Technical Information Serivce, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified		

Form DOT F 1700.7 (8-72)   Reproduction of completed page authorized

FOREWORD

One of the most remarkable feats of large-scale staffing for a highly techni-
cal occupation was the replacement of approximately 11,000 air traffic con-
trollers by the Office of Personnel Management and the Federal Aviation
Administration (FAA) following the PATCO strike in August 1981. This was
enabled, in part, by extensive research regarding air traffic controller
performance requirements and by the new Air Traffic Controller Selection Test
Battery developed by FAA research scientists. When adopted for use in October
1981, this controller selection battery reduced attrition in controller
training from 57 percent to 29 percent.

A major part of this report consists of an integrated presentation of the
significant research efforts that resulted in the adoption of the new selec-
tion test battery. The final section of the report describes the new National
Airspace System plan, announced by the FAA's Administrator in December 1981.
It analyzes the implications of this new automation program for the future of
the controller job and for selection of controllers in the new system. The
analysis presented suggests that much coordination may yet be needed between
systems engineers, who are designing the hardware and computer software, and
human factors specialists, who are concerned with the compatibility of the
system and the human operator, before a viable system is finalized.

This report not only brings together in one integrated volume the previously
scattered and largely unpublished results of a successful research activity of
the utmost importance to American aviation, it also sets the record straight
on many widespread misconceptions about air traffic controllers--such as
stress and the impact of prior aviation experience, age, sex, and education on
controller job performance. It presents a history, an up-to-date account of
the current situation, and a view of the future.

H. L. Reighard, M.D.
Federal Air Surgeon

DTIC
ELECTE
NOV 1 4 1984
B

Accession For

NTIS GRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	PER CALL JC

By

Distribution/

Availability Codes

| Dist | Avail and/or Special |

A-1

Herbert Ozur (Retired)   Research Psychologist, Applied Studies Group,
                            Office of Personnel Management, Executive Office
                            of the President, Washington, D. C.

Evan W. Pickrel, Ph.D.   Manager, Biomedical and Behavioral Sciences
                            Division, Office of Aviation Medicine, Federal
                            Aviation Administration, Washington, D.C.

Donald B. Rock   Director, Office of Personnel and Training,
                            Federal Aviation Administration, Washington, D. C.

S. B. Sells, Ph.D.   Research Professor and Director, Institute of
                            Bahavioral Research, Texas Christian University

Joseph A. Tucker, Ph.D.   Professor of Instructional Technology, The
                            Catholic University of America (Retired).
                            Consultant, Federal Aviation Administration,
                            Washington, D. C.

Allan D. Van Deventer, Ph.D. Supervisor, Selection and Validation Unit,
                            Aviation Psychology Laboratory, Federal Aviation
                            Administration Civil Aeromedical Institute,
                            Oklahoma City, OK.

CONTRIBUTORS

Neal A. Blake	Deputy Associate Administrator for Engineering and Development, Federal Aviation Administration, Washington, D. C.
James O. Boone, Ph.D.	Supervisor, Systems Analysis Research Unit, Aviation Psychology Laboratory, Federal Aviation Administration Civil Aeromedical Institute, Oklahoma City, OK. Now National Manager, Human Relations, Federal Aviation Administration, Washington, D. C.
Leland D. Brokaw, Ph.D.	Technical Director, Personnel Research Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas (Retired)
William E. Collins, Ph.D.	Supervisor, Aviation Psychology Laboratory, Federal Aviation Administration, Civil Aeromedical Institute, Oklahoma City, OK.
Joseph G. Colmen, Ph.D.	President, Education and Public Affairs, Bethesda, MD.
John J. Convey, Ph.D.	Associate Professor and Chairman, Educational Psychology Program, The Catholic University of America
John T. Dailey, Ph.D.	Manager, Biomedical and Behavioral Sciences Division, Office of Aviation Medicine, Federal Aviation Administration, Washington, D. C. (Retired). President, Allington Corporation, Remington, VA.
Oakie Domenech, Ed.D.	Assistant Psychologist, Education and Public Affairs, Bethesda, MD.
Jack M. Greener, Ph.D.	Assistant Professor and Head, Industrial Psychology Research, Institute of Behavioral Research, Texas Christian University
Thomas F. Hilton, Ph.D.	Research Psychologist, Naval Health Research Center, San Diego, CA. Formerly Graduate Research Assistant, Institute of Behavioral Research, Texas Christian University
Joanne Marshall Mies, Ph.D.	Associate Psychologist, Education and Public Affairs, Bethesda, MD. Now at Advanced Resources Research Organization, Washington, D. C.
Ann Milne, Ph.D.	Associate Psychologist, Education and Public Affairs, Bethesda, MD. Now at Applied Urbanetics, Inc., Washington, D. C.

PART III (Continued)

PART IV
RESEARCH LEADING TO THE 1981 ATC SELECTION BATTERY

TABLE OF CONTENTS

Chapter 1

INTRODUCTION AND OVERVIEW

S. B. Sells and Evan W. Pickrel

Two events around the end of 1981 marked the culmination of developments, initiated considerably earlier, that must be regarded as major landmarks in the history of the National Airspace System (NAS) of the United States. These were the adoption of the new selection test battery for ATC applicants, in October 1981, and the announcement of the new Airspace System Plan, in December 1981 (Federal Aviation Administration, 1981). The new test battery is the outcome of a test development and validation research program of a caliber comparable with that of the pilot selection program of the United States Army Air Forces in World War II, and has already reduced attrition in ATC training from 57% to 29%. One of the major purposes of this book is to present the relevant research on the ATC test battery, previously widely scattered in published and unpublished technical reports and memoranda, in a single, systematic document.

The new NAS Plan will not only replace aging and obsolescent equipment with modern, state-of-the-art, solid-state, computers and communications equipment, for greatly increased safety and efficiency, but will also introduce higher levels of automation, modernize the flight service station network to improve the dissemination of flight and weather data to pilots, and deploy new radar, communications, and airport landing systems to further enhance safety and provide more efficient traffic flow. By the year 2000, the system is expected to provide automated decision-making. One of the salient effects of the new system, that requires the most careful study, will involve changes in the role of the air traffic controller and possibly new requirements for controllers at some point in the transition. A second purpose of this book is to review the new NAS Plan from the vantage of controller selection and to begin to identify changes in the role and functions of the controller in the future, as the system evolves from its present (1982) form. The Plan is to introduce changes in an evolutionary fashion and projections concerning controller functions must take account of the expected status of the system at different stages of the replacement process.

The new test battery for controller selection must be regarded as a major advance in the application of scientific, actuarial methods to the pre-employment selection of personnel for a highly skilled technical specialty. It involved both innovation in test content, compared to the approaches that had become standard since World War II, and comprehensive concern with virtually every detail of a complex, sensitive area. The story of this outstanding program, which is reported in the first 23 chapters (Parts I through IV), is both a research document of major importance and a reference document for professionals and students in the field of personnel selection.

At almost the same time that the new test battery became operational, the announcement of the new NAS Plan gave warning that changes would be required in the selection program, perhaps within the next 10 years, depending

on the rate at which new computers, radar, and communications systems come
on line and pre-empt the customary functions of the controllers, and the ex-
tent to which the resulting changes levy new requirements for those who will
perform the new functions. These issues are addressed in Part V, which sum-
marizes the projected developments, examines the changes expected in the job
of the controller, and suggests new research and development to bring the
selection procedures in line with requirements. Much of this analysis is
necessarily speculative. However, the first step in every new research enter-
prise must be concerned with conceptualization, formation of hypotheses, and
definition of alternatives, and at the outset must be considered speculative.

ISSUES IN PRE-EMPLOYMENT SELECTION RESEARCH

The purposes of pre-employment selection for any job, but particularly
for those such as airline pilot and air traffic controller, that are highly
technical and involve special skills, are to identify, within a population of
applicants, the subset that is most likely to complete technical training and
perform the job successfully at the entry level, and to advance to higher
levels of skill and responsibility, with experience on the job. The appli-
cation of scientific methods to selection, based on field validation of
objectively scored tests and other predictors, against reliable measures of
performance of the job, was demonstrated on a large scale in programs to
select pilots and other air crew members by the Aviation Psychology Program
of the United States Army Air Forces in World War II. More extensive discus-
sions of the issues addressed in this section can be found in Guion (1965),
Magnusson (1966) and other textbooks.

Actuarial selection, as this is generally called, stands in contrast to
clinical selection, which is based most commonly on interviews by personnel
specialists, supervisors, or psychologists. The critical differences between
the two approaches are that: (1) actuarial selection is based on probabili-
ties generalized from empirical experience, using a scientific research design,
while clinical selection is more intuitive, depending on the judgment and ex-
pertise of the interviewer, for which probabilities are not usually calculated;
and (2) in actuarial selection, probabilities are used for groups or "batches"
of applicants who have attained particular scores, but not for individual
applicants, whereas clinical selection attempts to form a judgment concerning
each applicant. The history of pre-employment selection favors the accuracy
of actuarial selection for occupations involving large numbers of employees,
by an overwhelming margin, but only in those instances where the methods have
been appropriately employed.

Actuarial methods are very demanding and require research designs for
validation in which special attention is required to the samples of subjects
tested, the selection of predictor measures, the conditions of data collection,
the criteria used as measures of job success, and the type of analysis per-
formed. Each of these components is critical, as discussed briefly below.

Subject samples. Because scores based on selection tests and other pre-
dictors have probabilistic implications, prediction equations must be based on
large samples of applicants (or incumbents, as discussed later). Although the
definition of "large" may vary with the situation, sample size should generally
exceed several hundred and several thousand would be preferred. In order to

generalize experimental data to operational conditions, it is desirable that the composition of the test population be comparable to that of the target population with respect to status at the time of testing, and proportions by age, sex, race-ethnic group, educational background, and other variables deemed to be relevant.

Selection of predictor measures. Since the utility of a selection instrument (e.g. test) or battery of instruments depends mainly on its correlation with an acceptable criterion of job success (that is, its validity), and since testing and validation of experimental predictors is expensive and time-consuming, it is desirable to start out with a set of candidate predictors that are likely to correlate well with the chosen criterion. This is a most important step that should not be taken without insightful job analysis, to identify the critical aptitudes, background, knowledge, skills, and other human attributes involved in learning to do the job and performing it success-fully. This implies intimate knowledge of the target job with respect to the environment, equipment, procedures, and organizational structure involved, the training and prior experience required, the criteria for distinguishing effec-tive from ineffective performers, and the characteristics of effective and ineffective performers. Whenever any of these parameters changes, it is im-portant to assess the effects of such change on the person requirements.

The predictors that are finally selected from among all the candidate measures tried out experimentally will be those that correlate most consis-tently and most highly with the criterion and that do so independently of other predictors. It should be understood that reasonably high correlation between a predictor and a criterion implies that individuals who score high on that predictor also score high on the criterion (of job performance), and vice-versa. This is the acid test of predictive validity for any test or person characteristic assumed to be relevant to job performance.

All predictors that pass this test are relevant, but some that are con-sidered relevant may not predict. Failure to predict may reflect: (1) that the proposed predictor is not related to the criterion (not relevant), or (2) that virtually all candidates have similar scores (there is no variance) on the predictor (e.g., if basic literacy were included as a test for air traffic controllers), or (3) that the factor measured by the predictor is either not represented in the criterion or unimportant in the criterion as measured (this is frequently the case with such personality variables as social adaptability and dependability). Of course, predictor tests, parti-cularly those that measure these and other personality variables, must also be assumed to have demonstrated validity as measures of what they purport to measure, before they are accepted for try-out.

Predictive validity, as represented by the correlation of experimental predictors with criteria, is an objective, empirical test of the acceptability of proposed predictors in selection research. If "expert" opinion is valid, it can be confirmed by such a test. If it fails the test, there is a data-based reason to reject it. In dealing with human behavior, almost anyone can be an expert, but validation research is the only dependable test.

In selection research in general, the best results have been obtained with tests of cognitive and motor skill aptitudes and abilities and with tests

3.

of technical knowledge, which are most commonly well-represented in the criterion measures. On the other hand, results with personality and temperament measures have been at best marginally successful, partly because of limitations of the personality measures commonly employed, and partly because of the way that criterion measures are defined. In addition, an important requirement for all measures used, both predictor and criterion, is reliability, or consistency, which places an upper limit on any correlations that may be obtained. All too often, the graveyard of failed selection studies is marked by unreliable predictors or criteria, or both.

Conditions of data collection. The term prediction is frequently used as a goal of correlation, in the sense that a high correlation enables the prediction of the criterion based on knowledge of the (predictor) test scores (by means of a regression equation). Prediction is also used in another meaning, to distinguish between two types of study design. The first is the Predictive Study, in which the subjects are job applicants and the criteria are determined at a future time (at the end of training or on-the-job after completion of training) for those who are hired. The other is the Concurrent Study, in which the subjects are all employees (in training or further along on the job) and the criteria are determined concurrently (or sometimes even retrospectively).

There are advantages and disadvantages to both approaches. The predictive design captures the competitive motivation of job applicants, to do well on the tests, and is therefore realistic. However, it requires time delay for the criterion measures to mature and this may be administratively objectionable. The concurrent design utilizes subjects who have already been hired and their motivation in completing the tests might adversely affect the test results. However, it produces results more quickly and often appeals to administrators because of the convenience of the procedure.

Some investigators have arranged for tests taken in concurrent studies to be treated confidentially, so that individual scores are withheld from the employer. This is a desirable practice since it assures the employees tested that their test scores will not affect their employment status.

Both designs are affected by the fact that since their data are based only on employees, who are presumably qualified for the job, the range of test scores and criterion measures is restricted, compared to the range that would be observed in a population of applicants. Restriction of range has the effect statistically of reducing the magnitude of correlations for any data set, and therefore of underestimating the validity coefficients computed by employee samples. The studies reported in this book reflect both types of design and include procedures developed to make corrections for restriction of range, in order to estimate the true correlations (validity coefficients).

Criterion measures. Perhaps the greatest limitation observed in selection research, over the years, has been in the shortcomings of measures adopted as criteria of job success, with respect both to their reliability and also to the extent to which the measures used in various studies have truly represented actual job performance. Obviously, assessment of job performance is difficult; for example, questions such as how well does an

4.

airline captain perform his job? or an EnRoute or Terminal tower controller? require complex decisions concerning behaviors to be included and design of tailored instruments to measure them. Without going into detail at this point (this topic is addressed in Part III), it is common knowledge that the most reliable, quantitative, and comprehensive criterion measures for selection purposes have involved measures representing assessment of performance during training, such ás .pass-fail of the training program, course test scores, and grades, and comprehensive final assessments. For on-the-job performance, global ratings by supervisors are generally believed to be less reliable than differential, task-oriented, behavior-based supervisor ratings, but both have been used successfully in carefully structured situations; trait-oriented rating forms (e.g., honest, reliable, motivated, etc.) have been found to be generally worthless.

The criterion measures employed in the research reported throughout this book included both during-training and on-the-job assessment, and the rationales, methods of data collection, and in most cases, the reliability of the measures used, are reported in the text.

Analytic methods. The correlation of a predictor measure with the criterion is a "raw" validity coefficient. The extent to which this figure can be generalized for future operational use of the predictor as a selection instrument depends on the representativeness of the sample with respect to the target population; the sample size, which affects the stability of the data; the conditions of test administration and data collection with respect to the status of the persons tested as applicants or as employees; and confidentiality. Some of these issues can be examined statistically; for example it is possible to test the representativeness of sample data; to analyze variations among subgroups defined by age, sex, race-ethnicity, or other factors; to calculate reliability estimates and standard errors of estimate; and to estimate "true values," making adjustments for unreliability of measures and restriction of range.

The correlation between a single predictor and a criterion is designated by the symbol "r," while the multiple correlation between a set (or battery) of predictors and a criterion is designated by the symbol "R." The prediction equation calculated from the data used in computation of r is called a regression equation, and from R, a multiple regression equation. The weights assigned to predictors in a multiple regression equation, indicating their differential contribution to the total, composite prediction score, are called regression weights, and these may be expressed in standard score form (beta weights) or raw score form (b weights).

A procedure that has become common practice in validation research is that of testing the stability of miltiple regression results by cross-validation on one or more independent samples. Failure to replicate results by cross-validation throws doubt on the initial results, while successful cross-validation increases the likelihood that the results will generalize by the product of the probabilities obtained independently in the samples investigated.

Interpretation of the magnitude of r or R, which affects the accuracy of predictions based on the relationship denoted, is based on the coefficient of determination, obtained by squaring r or R. The squared correlation (r^2 or R^2)

5.

indicates the percentage of common variance, or the extent to which the predictor(s) and criterion reflect common factors. Figure 1 shows that r^2 (or R^2) is initially lower and then increases by progressively larger steps as r(or R) increases in uniform steps from zero to unity, illustrating that the magnitude of the relationship changes with equal increments of the correlation coefficient. Thus, a correlation of .40, which is in the higher range of correlations usually obtained in industrial pre-employment selection research, accounts for only 16% of variance in common with the criterion, while a correlation of .20, which appears half as great, accounts for only 4%. It should be noted that correlations in the range of .40 to .60 can provide extremely effective selection, when compared with alternatives that correlate with the same criteria at levels below .20.

Significance and causality. A universal condition of human measurement is that of error, which is reflected in test scores as a result of a myriad of causes, such as ambiguities and other shortcomings of tests, distractions or other aspects of the testing situation, variations in individual efficiency, attitude, and well-being at the time of testing, errors in scoring, transcribing, and processing of data, and many others. Consequently, psychometricians assume that an actual, obtained score, as well as statistics computed, such as means, is composed of the true score plus error. While every effort is made to exclude error that can be controlled by the examiner (by maximizing test reliability and validity, standardizing testing and procedures, and checking all data carefully), probability statistics are employed to estimate error and to guide interpretation and adjustments to compensate, where possible. For example, the reliability of a test can usually be increased by increasing the number of items, assuming that a pool of relevant, unambiguous, and properly constructed items is available initially, and the statistical significance of a correlation coefficient can be increased by increasing the size of the sample of persons tested, assuming that the age, sex, ethnic, and other background mix remains the same, if additional subjects are available.

Significance tests are usually evaluated in terms of probability. For example, the standard error of a correlation coefficient is used to calculate the probability that the obtained correlation differs significantly from zero, and this is conventionally expressed as a p (probability) value. The borderline value $p < .05$, generally accepted as indicating significance, means that the chances are less than 5 in 100 that a correlation of the magnitude obtained would occur by chance alone (between the variables involved) in an infinite series of replications of the data. Similar interpretation applies for $p < .01$ (one chance in 100) and other values that may be calculated.

Since the standard error (and p value) is a function of sample size, a distinction must be made between statistical significance, which refers to the probability of approximating the true correlation for the population from which the study sample was drawn, and practical significance, which refers to the power of prediction afforded by the correlation obtained, regardless of statistical significance. A correlation of .10, obtained for a sample of 10,000 applicants, would be highly significant ($p < .001$) statistically, but would not afford significant predictive power.

6.
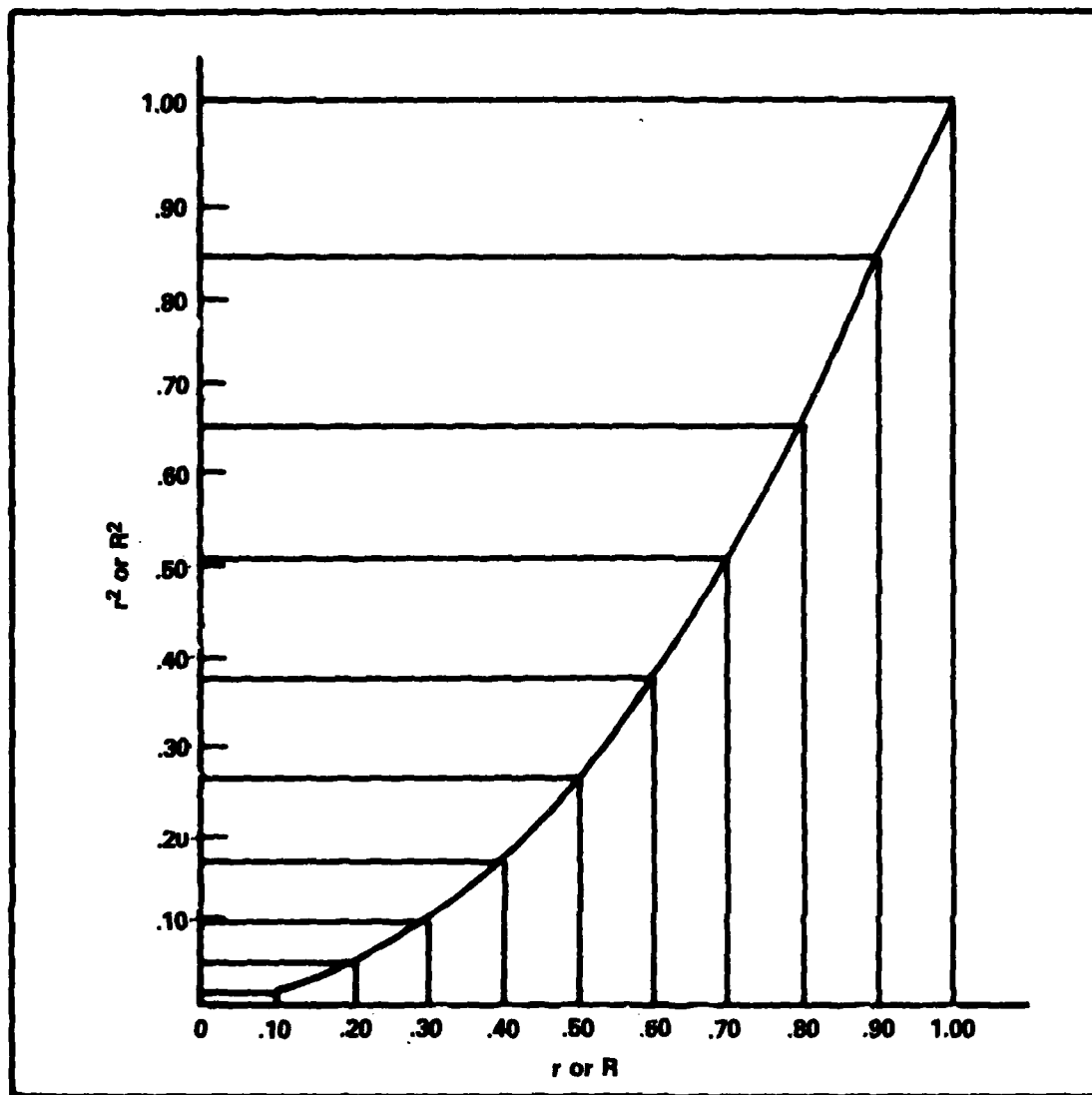
Figure 1.
*The Relation Between the Coefficient of Correlation (r) or
multiple correlation (R) and the coefficient of determination (r² or R²).*

7.

Another important distinction to be understood is between correlation, which describes the degree of covariation between a pair of variables, and correlation, which describes the degree of covariation between a pair of variables, and causality, which implies that the relationship between them is such that the occurrence of one results in a predictable effect with respect to the other. Often, the correlation between two variables may be the result of their common relation to a third variable, as for example, height and scores on a school achievement test correlate fairly well (around .60) because of their common relation to age. As a general rule, one should make no inferences concerning causality from correlation data alone.

Utility of proposed tests. After the predictive validity of a proposed test or test battery is determined, there still remains the task of analyzing its utility for operational use. This involves a number of considerations, including (1) the development of composite predictor scores, (2) the gain, over existing selection procedures, in the percentage of successful candidates, at various levels of predictor score (cutting scores), which can be demonstrated by means of some form of expectancy table, (3) the consistency of various cutting scores for subgroups of the applicant population (men, women, race-ethnic groups, etc.), (4) the yield in successful cases in relation to testing time required; and (5) assessment of the consequences of shortening the battery (if necessary), in terms of yield. Related considerations, such as the feasibility of including tests that require special apparatus for administration, and the inclusion of items that may for any reason be controversial, should be addressed in the preliminary stages of the research. Finally, problems of security of tests and scoring keys must be addressed in order to prevent compromise, should control of vital information be lost. This usually implies the necessity of constructing sufficient alternate forms of tests to enable replacement immediately should a need arise.

Fairness. In a diverse, multi-ethnic population, such as that of the United States, and also in bilingual populations, as in Canada, it is possible to give an advantage to one group over others that may result in significant differences in test scores, by choosing test items that afford differential facility in responding, because of differential linguistic, cultural, educational, and economic background. Indeed, this problem even applies to sex groups, since male and female developmental experience, and hence knowledge, skills, and physical capabilities are in many ways a function of prescribed sex roles.

Concern with this general problem arose in the United States in the 1930's, when it was claimed that mean differences, often as high as 15% to 20% between white and black school children on intelligence and other verbal tests were attributable to built-in cultural bias in tests, designed mainly by middle and upper-middle class white psychologists, rather than to biological differences between the "races." The controversy on this issue has raged for over 50 years and is not yet resolved. However, efforts have been made in educational testing to devise "culture fair" and "culture free" tests, but with extremely modest success.

In the area of pre-employment testing, the issue of fairness is complicated. Although it is considered desirable to give every applicant an equal opportunity to be selected for a job for which he or she applies, without

8.

penalty for race, color, sex, age, or other personal attribute, it is also
imperative that persons selected, particularly for such critical jobs as air-
craft pilot and air traffic controller, be capable of performing the job suc-
cessfully. Both of these issues were recognized by the Equal Employment
Opportunity Commission, the Civil Service Commission (now the Office of
Personnel Management), and other cognizant federal agencies, and the Uniform
Guidelines in Employee Selection Procedures, adopted for federal employment
in 1978, provide what is generally regarded as an equitable solution. This
is explained in Chapter 22 in relation to the new controller selection test
battery adopted in October 1981, and essentially prescribes that if adverse
impact for any sex, age, or racial minority groups is found for any proposed
selection procedure, then a positive demonstration of predictive validity
(job relevance) and fairness (in application of regression results) must
be made.

WHAT AIR TRAFFIC CONTROLLERS DO
AND THE SYSTEMS IN WHICH THEY WORK

The purpose of Air Traffic Control (ATC) was described in a report on
Air Traffic Control Specialist (ATCS) training (Henry, Kamrass, Orlansky,
Rowan, String, & Reichenbach, 1975), as follows:

> "The purpose of air traffic control (ATC) is to ensure safe
> and efficient movement of aircraft. Control of aircraft in the
> National Airspace System (NAS) is done from the ground, and it is
> designed to keep aircraft separated from each other and to expedite
> the flow of traffic. Control facilities develop information from
> a variety of sources, including long-range surveillance radars,
> local airport radars, adjacent control areas, and by direct vision
> from towers. Air traffic control is exercised at terminals and
> between terminals. Control between terminals is called en route
> control. En route control in the continental United States is
> distributed among 20 air route traffic control centers (ARTCCs)
> for aircraft operating on instrument flight rules (IFR). Terminal
> control facilities can be divided into those capable of handling
> traffic operating on IFR and those that can handle only aircraft
> operating under visual flight rules (VFR). Controllers are usually
> classified by the kind of facility in which they operate, i.e., en
> route, IFR terminal, and VFR terminal. Training programs for
> controllers are designed for each of these specialties" (p. 7).

Air traffic controlling involves many human tasks that have been clus-
tered into a set of jobs that comprise the ATCS occupation. In broad terms,
the skills required to perform these jobs include problem solving, decision
making, information processing, oral communication, coordination, and equip-
ment operation. The knowledge requirements are extensive and relate to
regulations, operating procedures, air navigation, piloting, weather, and
sector mapping. Since the critical skills are cognitive, selection proce-
dures emphasize assessment of cognitive facility.

The jobs of the ATCS can be described in terms of these categories,
located at Terminal (Tower), Air Route Traffic Control Centers - ARTCCs

9.

(EnRoute Centers) and Flight Service Stations (Stations). Flight Service Station (FSS) specialists provide advisory services only, to general aviation pilots. The following discussion of these specialties includes excerpts from the description of the new NAS Plan (Federal Aviation Administration, 1981).

Terminal Air Traffic Control. The Terminal air traffic controller works with aircraft during takeoff and landing, using direct vision, radio communication, or radar to obtain information concerning the position and course of the aircraft, and communicating with pilots by radio.

"The three major types of facilities used in terminal air traffic control are the airport traffic control tower (ATCT), terminal radar approach control (TRACON), and terminal radar approach control in the tower cab (TRACAB). Located on airports, the ATCTs are the most common, as well as the most visible, of the three. Their purpose is to separate aircraft, sequence aircraft in the traffic pattern, expedite arrivals and departures, separate aircraft on the landing areas, and provide clearance and weather information to pilots.

"The second most common are the TRACONs that control airspace around airports with high-density traffic. TRACON controllers separate and sequence both arriving and departing flights. Normally, each TRACON is associated with one ATCT and located within the same building. However, a TRACON may be remotely located and may serve more than one ATCT. The third type, the TRACAB, serves a function similar to that of the TRACON. TRACABs are located within tower cabs at airports with low-traffic density.

"All terminal air traffic control facilities are equipped with radio communications to aircraft, telephone communications to air route traffic control centers and flight service stations, and have a variety of equipment for observing, detecting, receiving, and displaying weather information.

"Radios and telephones are major tools of the terminal controller.

"Terminal ground-to-air communications are conducted with VHF or UHF transmitters and receivers. Current ATC ground-to-air communications require many frequency changes as aircraft move through the system. Terminal ground-to-ground communications use a variety of old switching systems. These systems are leased and range in complexity from the systems used in many small ATCTs to Western Electric 301, and 301A systems used at TRACONs.

"The display of information on aircraft position in the traffic pattern, aircraft identity, and aircraft altitude greatly assists the controller. This information, obtained by surveillance radar, is required in areas of high traffic density. Radar information assists in separating aircraft, expediting the traffic flow and allowing separation minima to be greatly reduced from those in non-radar operations. Terminal radar information is provided by airport surveillance radar (ASR); i.e., primary radar and air traffic control beacon interrogators (ATCBI) (i.e., beacon).

"Radar information currently is processed and displayed by a variety of automation and display equipment depending on traffic density, required

10.

capabilities, and the display environment. Current automated processing and display equipment came into general use during the 1970s.

"Computers not only have relieved the controller of routine tasks but also give the controller information that assists in maintaining aircraft identification. In addition, information such as altitude and aircraft speed, which must be provided by the pilot in a non-automated environment, can be displayed by means of alphanumeric symbology.

"The TPX-42, which is the least sophisticated of the terminal automation systems, is a non-programmable, numeric beacon decoder system. That is, information from an aircraft's transponder is decoded and displayed for the controller in numeric form along with normal radar data. It provides aircraft identification and altitude information for suitably equipped aircraft. TPX-42 is used at lower activity facilities.

"The automated radar terminal system (ARTS II) is a programmable, non-tracking data processing system. Although it does not provide tracking, the ARTS II system does provide meaningful information to the controller, such as aircraft identification and altitude. Like TPX-42, the ARTS II derives its information from the aircraft's transponder. The ARTS II system may be interfaced with the ARTCC computer for the automatic exchange of information. The ARTS II is used for facilities having low-to-medium activity. At present, ARTS II has no unique software other than the basic operational program.

"The automated radar terminal system (ARTS III) is a programmable, beacon tracking system. Based on information from the aircraft's transponder, the ARTS III detects, tracks, and predicts the position of aircraft in the terminal area. The information is displayed on the controller's radar scope by means of computer-generated symbols and alphanumeric characters, along with normal radar data. The computer displays aircraft identification, altitude, ground speed, and flight plan data. In addition, the ARTS III is interfaced with the ARTCC computer, allowing the computers to pass information between facilities.

"As of 1982, ARTS III has been installed at medium-to-high activity terminal facilities and has provided two unique features, through programming. The first is the minimum safe altitude warning (MSAW). MSAW is a function of the ARTS III computer, that will alert the controller when a tracked aircraft, with altitude reporting capability, is below or is predicted by the computer to go below a predetermined minimum safe altitude. The second is conflict alert. Terminal conflict alert (CA) is a computer function that alerts the controller to the situation that aircraft are in close proximity and possibly require attention or action.

"The automated radar terminal system (ARTS IIIA) is an enhancement of the existing ARTS III system designed to meet increasing traffic demands in terminal airspace. The ARTS IIIA is capable of tracking radar targets as well as transponder-equipped aircraft. Thus, all aircraft that are within radar coverage of an ARTS IIIA facility are candidates for computer processing. Like ARTS III, ARTS IIIA interfaces with the ARTCC's computer and contains unique software features of MSAW and CA.

11.

"Flight data entry and printout (FDEP) is not, in and of itself, automated equipment. However, it does provide an automated means of printing terminal flight progress strips. FDEP equipment interfaces with the ARTCC's computer and is used extensively throughout the system to automatically pass flight plan information from ARTCCs to terminal approach control facilities and associated control towers.

"Bright radar indicator tower equipment (BRITE) is a supplemental radar display system designed for use in the bright light environment of the ATCT cab.

"The current terminal ATC system consists of over 400 ATCTs and nearly 200 TRACON/TRACABs" (Federal Aviation Administration, 1981).

EnRoute Center (ARTCC) Air Traffic Control. Controllers at Centers monitor some general aviation and all commercial aviation flights while they are en route to their destinations. They also issue instructions to pilots on proper altitudes and flight headings that they decide are necessary to maintain legal separation from other aircraft. They instruct pilots to ascend or descend to new altitudes or to change course to avoid severe weather or restricted flight areas.

"A typical center is responsible for more than 100,000 square miles of airspace and hundreds of miles of airways in the sky which are like electronic highways to pilots. A center's geographic area is uaually divided into 30 or more sectors, with a team of controllers responsible for each sector.

"There are 20 air route traffic control centers in the continental United States. There are five offshore centers located at Anchorage, Honolulu, San Juan, Panama, and Guam. The San Juan, Panama, and Guam centers are combined center radar approach controls (CERAPs).

"Another integral part of the en route system is the central flow control facility in Jacksonville, Florida. Central flow control serves as a focal point for evaluating and approving traffic flow redistribution, nationwide management of air traffic flow, and provides authority for initiating systemwide flow control. Central flow control, associated with the airport reservations office (AR), relieves congestion at the busiest airports. When associated with the central altitude reservation function (CARF), it supports military operations and provides coordination of other activities requiring airspace protection.

"These en route centers control all aircraft in the United States operating under instrument flight rules (IFR) and not under control of military or other facilities. They provide separation services, traffic advisories, and weather advisories. In addition, they track aircraft operating under distress. The FAA en route system is, of course, an integral part of this country's national defense system.

"The air traffic control system determines correct aircraft separation based on radar data input. It also provides visual flight rules (VFR) traffic advisories, and fixed route clearances based upon separation.

"Most flight data, after being processed, are now displayed on paper strips torn from flight strip printers. This is a mechanical system requiring manual coordination and input by the air traffic controllers. Handoffs of aircraft from one controller to another, together with the required frequency changes, are also done manually.

"The en route centers presently use 9020 computers developed in the 1960s. The three offshore control centers use en route automated radar tracking systems (EARTS) computers to perform similar but more limited radar and flight data processing" (Federal Aviation Administration, 1981).

Flight Service Station Specialists (FSSS). The FSS function is mostly advisory to the general aviation community (as distinguished from commercial air carriers). It includes the following specific services: accepting and closing flight plans, briefing pilots, en route communications with pilots flying VFR, assisting pilots in distress, disseminating aviation weather information, monitoring radio navigation stations, originating notices to airmen, working with search and rescue units to locate missing aircraft, and operating the national weather teletypewriter systems.

There are more than 300 FAA Flight Service Stations.

"At certain locations, flight service stations take weather observations, issue airport advisories, provide en route flight advisory service, and advise customs and immigration officials of transborder flights. The stations also have communications equipment for relaying information to towers and air traffic control centers and for various other emergency services.

"Of all the FSS services, none are more important to safety than those related to weather.

"The FAA aviation weather system collects weather information and distributes it to both pilots and agency operations personnel. Weather information is collected largely with electromechanical devices that give wind direction and velocity and measure cloud ceilings. Facsimile weather maps and low-speed teletype equipment also are used.

"FAA long-range air traffic radars also give two levels of contours to outline weather on en route radar displays for controllers and for center weather service unit (CWSU) meteorologists at the en route centers. Other aviation weather information comes from the National Weather Service, Aeronautical Radio, Inc. (ARINC), pilot reports, and observations made by agency personnel.

"To get weather information to pilots, the agency depends on telephones and radio voice broadcasts, including advisories made over VOR radio stations used for navigation. At some locations, voice recordings disseminate mass weather information. For pre-flight briefings and in-flight advisories, direct communications between the pilot and flight service station specialist are used" (Federal Aviation Administration, 1981).

Career path. Terminal and EnRoute air traffic controllers have a career program ladder that extends from the entry level (GS-7 for competitive

13.

appointees from the OPM register or GS-5 for noncompetitive appointees) to
GS-13, and GS-14 for Full Performance Level (FPL) journeyman radar control-
lers. Following initial centralized training at the FAA Academy at Oklahoma
City, the graduate works initially as a Developmental controller at the facil-
ity at which he or she was hired. Developmental controllers receive on-the-
job training and performance assessment and may progress through a series of
competency levels that terminate with achievement of FPL, journeyman status.

The FPL designation implies that an ATCS is fully certified to work at
any controller position at a facility. Developmental training may require
5 or more years to accomplish FPL status. Previously, unsatisfactory pro-
gression was a basis for termination of employment at any stage of develop-
mental training. However, as pointed out in Chapter 6, the system was
changed in 1981 after the PATCO strike. In the new personnel system radar
training is an advanced course at the Academy, for which developmental non-
radar controllers must qualify. Non-radar controllers who fail to qualify
for or to pass the radar course may now work out their careers in the non-
radar status. Those who do qualify for and complete radar training, and who
advance to FPL status, are expert professional specialists who have passed
through rigorous screening, training, and on-the-job developmental training
to qualify for this certification.

PROJECTED NATIONAL AIRSPACE SYSTEM CHANGES

In December, 1981, the Federal Aviation Administration announced the new
National Airspace System, which is intended to reach high levels of automation
around the year 2000. A detailed analysis of this plan, taking account of
its impact on the ATCS function, is presented by Neal A. Blake, in Chapter 24.
The conceptual approach and planned evolution of the Terminal, EnRoute, and
FSS portions of the planned system are explained in the following extracts
from the FAA December, 1981 NAS Plan report.

Terminal System

"The objectives of the terminal system improvement plan are to maintain
a very high level of safety, impose minimum constraints consistent with effi-
cient use of the system and at the same time minimize FAA operations costs.
This involves extended use of automation and consolidation of the number of
air traffic control facilities required. For the near term this means making
hardware and software improvements to existing ARTS automation systems, and
improvements to the facilities and voice communication switching which co-
exist with them. Over the intermediate term it involves replacement of the
automation equipment to provide capacity for traffic growth, higher system
reliability and the addition of safety and efficiency enhancements.

"A common family of computers and smart sector suites will eventually
be employed to improve air traffic productivity and provide significant cost
savings by reducing the multiple life-cycle costs for training, supply support,
engineering and software associated with TPX-42, ARTS II, ARTS III, and
ARTS IIIA. Computers provided to hub TRACONs will be subsets of those used
for the en route centers and will employ common software. Sector suites used
at hub TRACONs will be identical to those developed for the en route ARTCC
and have identical processing capability.

"A modular version sector suite will be developed for ATCT use. It will contain identical processors but have unique displays for use in the space-limited, high intensity light environment of the ATCT cab. Sector suites will draw on terminal, en route, and FSS data bases and will satisfy the traffic control requirements for radar flight position, intent and identify information, flight data information, weather data information, and flow planning information.

"In the long term, automation will be extended to create an integrated flow management system which will maximize airport capacity, smooth traffic flow and reduce aviation fuel consumption. Consolidation of terminal air traffic control facilities will be undertaken to increase controller and technician productivity. The 188 TRACON/TRACABs which currently exist will be consolidated into 30 newly established hub TRACONs or existing ARTCCs by the mid-1990s. The trend over the period will be to blend separate functions of terminal and en route air traffic control into 48 major ATC facilities and supporting ATCTs. Facility consolidation is expected to reduce the flight coordination between pilots and controllers and result in significant savings in equipment, personnel and operating costs. Upgrading automation and voice communication switching will be the cornerstones of facility consolidation."

Near Term - to 1985. Near term improvements (to 1985) will "focus on higher productivity among air traffic controllers and Airway Facilities maintenance technicians and control of communication costs. ARTS II machines will be upgraded. Software for conflict alert and MSAW safety improvements will be provided. Software to enhance conflict alert at ARTS III locations will be provided to reduce nuisance alarms. Enhancement of the controller training capabilities at ARTS III locations will be provided to expedite strike recovery.

"ARTS III memory will be expanded to provide capacity for traffic growth. The capability to hand off aircraft and transfer flight data automatically between ARTS II and ARTCC facilities will be added to approximately 50 locations to reduce controller workload. A similar ARTS-to-ARTS direct interface will be created. An automated ATC capability in the R-2508 restricted area of California will be created to prevent midair collision of civil and military aircraft.

"Radar remoting to satellite ATCTs will be provided to reduce collision risk. Vacuum tube BRITE I and II displays will be replaced to improve technician productivity. TRACON BRITEs will be eliminated to improve working conditions. An integrated communications switching system (ICSS) will be purchased on a competitive basis to decrease leased communications costs. Power conditioning systems will be provided to ARTS III locations to improve reliability. Research and development will start on sector suite and computer modernization with a host machine.

"Research and development of automated airport advisory services are underway to determine the operational performance of a system targeted to supplement or replace VFR tower service. Operational test of concepts to improve airport capacity will be conducted. Operational test of a runway configuration management system will be conducted at Chicago O'Hare. Research

15.

on integrated flow management is underway. Sustaining ATCT and TRACON modernization and relocations will continue. Construction of hub terminals will begin. .

Intermediate Term - 1990. "During the intermediate term (1990), TPX-42s will be replaced and software for conflict alert and MSAW will be provided. Terminal hub consolidation and terminal-to-center consolidation will be well underway. Sector suites, tower sector suites, voice switching and control systems, and subsets of en route host computers will be combined to allow operations personnel to handle increased traffic without a proportional increase in staffing. Tower communications switching systems will be provided to ATCTs which have not received ICSS in the near term to (1985). The large geographic areas of the new 30 hub TRACONs and the absorption of TRACON facilities into the ARTCCs will increase the similarity of terminal and en route systems. Improved weather product utilization will result from the computer and sector suite. ATCT sustaining relocations and modernization will continue along with research into integrated flow management.

Long term (to 2000). "Over the long term (2000), the FAA plans to complete terminal hub consolidation and terminal-to-center consolidation. Computer software will be implemented for integrated flow management products. Data link coverage will be extended down to 6,000 feet MSL and will be utilized to provide weather and clearances to aircraft.

EnRoute System

"Despite their capabilities, the current 9020 computer systems are not expected to prove adequate for handling the projected growth in aviation traffic beyond the late 1980s.

"Major enhancements to the operational software are not possible because of capacity limits. For instance, the system is incapable of ensuring 100 percent functional reliability or availability necessary for upgrading automation. These higher levels of automation are being developed to further reduce operational costs, improve safety, and provide fuel savings for aircraft users.

"In addition to its high hardware and software operating and maintenance costs, the present en route system is labor-intensive. Moreover, equipment manufacturers cannot continue to support the system indefinitely, regardless of the cost.

"FAA plans call for replacement of the central computer complex 9020 computers in the mid-1980s with a host computer. (A host computer is a replacement computer which uses existing software from another computer system.) This host computer will be capable of running the present 9020 software package with minimal modifications. This will provide immediate relief for capacity restrictions from anticipated air traffic growth.

"The new computer system will be designed for approximately 100 percent functional availability and reliability of en route services. Moreover, the new computers, software programs and displays, now under development, will be capable of providing both en route and terminal services. This will enable the agency to consolidate and reduce the number of facilities needed.

16.

Significant savings in manpower, rents, utilities, and energy costs will be realized. Use of the same system in en route and terminal facilities will also eliminate or considerably diminish the present, somewhat arbitrary, demarcation of services and thereby reduce operational overhead.

"At the same time, the agency will begin action to procure new sector suites which include distributed processing minicomputers capable of handling the functions now resident in the display channel computers. This contract will also provide for development of replacement software programs partitioned to run in the host computers and sector suites. Major operations requiring centralized processing will be accomplished in the host, with all remaining functions performed within the individual sector suites. This will minimize the impact of most failures to a single sector, with sufficient back-up capability to provide 100 percent functional reliability. It will also minimize the adverse impact of a major technical and operational transition. Additional safety and productivity functions will be included in the new software.

"The sector suite will consist of three displays providing situation display information, such as: radar/beacon coupled with the introduction of real-time weather, electronic tabular display (eliminating the need for the manual flight strip processing), and a third display for planning information and advanced functions associated with AERA. Connection to various universal busses will allow the sector suites access to all necessary information shared with other systems.

"The capability of the system to accomplish processing of both en route and terminal input and functions allows for the consolidation of terminal facilities into the centers and the use of the same or a downsized system in the terminal environment for consolidation of terminal radar approach control (TRACONs). Standardization, flexibility, and expandability of the system's design will allow more universal application, significantly reduce total operations costs, and facilitate accommodations of future requirements and enhancements.

"Integration of a voice switching and control system into this environment will provide further productivity gains through automatic switching of communications. This capability will allow resectoring consistent with demand requirements. Implementation will reduce leased line and equipment costs and permit eventual integration of voice and data communications over a satellite-based system. This will even further reduce transmission costs.

"During this period, the agency's replacement of the central flow control computer complex system at Jacksonville will improve central flow control and allow expansion to cover more of the system. The long-range goal will be its coupling with en route and terminal programs for a total national flow control concept. This will be more predictive and include consideration of essentially all of the National Airspace System.

"In summary, FAA's present en route plans are designed to replace current air traffic control systems to meet future needs of greater capacity and reliability. With higher levels of automation, present en route plans will improve both safety and productivity. The new system will be capable of providing both en route and terminal services, thereby enabling the agency to consolidate and reduce the total number of facilities needed to do the job.

"Near Term (to 1985). Over the near term (to 1985), the FAA will de-
velop a replacement for its 9020 computers with a new host computer capable
of running slightly modified 9020 software programs. The new host computer
is slated for complete development in 1985. It will provide the additional
capacity needed to handle the growth in air traffic expected beyond the mid-
1980s.

"During this same period, agency plans call for more automated interfaces
between en route and terminal facilities providing better coordination of
operations.

"Near-term agency plans also call for improving weather information by
combining center weather processing capability with weather information from
flight service stations. Replacement of mechanical, labor-intensive flight
strip printers (FSP) and flight data entry and printout (FDEP) equipment with
more reliable and faster electronic equipment is expected to improve produc-
tivity. Near-term software improvements will be aimed at improving airspace
efficiency, safety, and reliability.

"The agency plans to implement new en route metering for high-altitude
traffic. This should also cut back on delays and produce optimized fuel
descent profiles. Conflict alert will be expanded to include detection of
aircraft equipped with Mode C transponders when they intrude into controlled
airspace. Additional software improvements will include development of con-
flict resolution logic that will assist controllers in preventing violations
of separation standards. Other planned software changes will improve the
detection of hardware element errors and at the same time configure operating
elements into a working system.

"Similarly, the software for the en route automated radar tracking sys-
tems (EARTS) will be improved over the near term to provide conflict alert
and minimum safe altitude warning systems. Among the first improvements to
oceanic control systems will be automated conflict probes along with logic
needed to decide on the most fuel-efficient routing.

"Capability of the backup direct access radar channel (DARC) will be
expanded to provide full data information tags and individual display
switching. This will eliminate the need to convert the displays to a hori-
zontal position, to use shrimpboats, and to switch six displays at a time to
the backup channel, as is now the case.

"Successful completion of these near-term programs will enable the air
traffic control system to handle the growth in traffic forecast for this
period. There will also be benefits to the users in the forms of:
(1) greater efficiency through en route metering and direct route processing,
and (2) improved safety and airspace utilization via better weather information.

"During this period, the 9020 computer at the Jacksonville, Florida,
central flow control computer complex will be replaced by an interim IBM 4341
computer and the complex will be relocated to Leesburg, Virginia, for better
coordination with FAA's Washington headquarters. The manual altitude reser-
vation facility (CARF) and airport reservations office (ARO) functions will
also be automated at that time. In the near term, standard remote control

18.

equipment will be developed and implementation of the 2,400 tone control channels will begin.

"Currently, there are 25 air route traffic control centers (including three CERAPs). By 1985, consolidation will reduce that number to 21 air route traffic control centers (including one CERAP).

"Intermediate Term (to 1990). During this period the display channel processors and displays will be replaced by sector suites, including software development of redistributed processing between the host computer and sector suite processors. This will also allow for the integration of terminal functions within the ARTCC and serve as a basis for future levels of automation.

"Mode S will provide a major system enhancement by 1990. Mode S will furnish improved surveillance, especially in high-density terminal areas. In addition, data link provides rapid dissemination of control and other information to specific aircraft.

"These activities will result in enhanced safety levels through the use of IFR/VFR conflict alert and conflict resolution advisories. Greater government and user efficiency will be realized through: improved flight planning using sector suite facilities, reduced en route separation using Mode S with monopulse, expanded use of area navigation, and oceanic probes with displays.

"A complete voice switching and control system, which complements the capabilities of the new sector suite, will be developed and implemented during the intermediate period. This will allow dynamic reconfiguration of sectors, further reducing workload.

"The central flow control function will be significantly upgraded with the introduction of a new computer system. Along with software to enhance its ability to project and estimate NAS congestion and delay levels, it will evaluate alternate flow management strategies based on: arrival/departure messages, capacity, estimates from major terminals, data input from the flight service data processing system (FSDPS) on VFR flight plans, and input from the aviation weather processor (AWP) for national weather information.

"Sixteen ARTCCs in the continental United States, two offshore ARTCCs, and no CERAPs will exist by 1990. Center building and plant modernization will be concentrated in the 16 continental centers that will remain after the intermediate period.

"Long Term (to 2000). Activities early in this time period will provide the capacity and informational framework for higher levels of automation. Tools for operator use of the system at higher levels of automation will be installed. During this period, the AERA functions will be implemented incrementally in the following sequences: (1) direct fuel-efficient routing, (2) flow planning and traffic management, (3) strategic clearance delivery, and (4) full tactical control. The central flow control function will have been further upgraded to work with AERA and integrated flow management (IFM) functions. Direct user access capability will also be provided for the flying public.

19.

"These activities will result in higher levels of safety and efficiency through the use of automated conflict probe and resolution, systemwide direct routing, and the capability to operate a sector with one person.

"Services provided by this country's future en route system will be significantly better and more cost-effective; system reliability will be much higher. The increased use of automation will significantly improve both safety and efficiency. Introduction of more advanced automation functions into the system will result in ever greater fuel savings.

Flight Service Stations

"Flight services will be improved for pilots by giving them direct access to: weather information; flight delay information, both in the air and on the ground; and flight plan filing. Aviation weather service will be improved in quality and timeliness.

"Automation and improvement of flight service stations and related aviation weather systems will allow consolidation of facilities, reducing operating costs significantly. Weather radar and current weather information will be provided to en route and terminal controllers.

Near Term (to 1985). Procurement of new computer systems for the automation of flight service stations is already underway. "FSS specialists will be provided with more timely national weather data at their display consoles. Automation will greatly accelerate retrieval of weather information along flight routes and the entry of flight plans.

"Also during the near term, limited direct access for pilots will be provided by the interim computer-generated voice response system. This will permit pilots to retrieve information from an automated weather data base via telephone. Replacement of low-speed teletype with data terminal equipment also will begin. Development of improved weather sensors and products will continue.

"Automated weather sensors will provide current airport weather information directly to pilots and to the agency aviation weather collection system. Satellite weather photos will be available at certain locations over facsimile recorders, displaying the location of cloud cover and weather systems.

"Low-level wind shear equipment will be added at more airports to detect hazardous wind conditions along the final approach. Six levels of weather radar contouring, outlining storms, will be on television displays for en route meteorologists and the automated flight service station specialists.

"Intermediate Term (to 1990). Major improvements to flight service automation will allow direct pilot access to weather and system delay data bases by telephone or through remote terminals. Flight movement data processing will be improved. Computer-aided direction finder triangulation will expedite locating and assisting lost aircraft. Flight service stations will be consolidated. Because they are covering larger areas, the remaining 61 automated facilities will be expanded. New communications switching systems will improve pilot access and expedite coordination with flight service specialists. The replacement of low-speed teletype equipment will be completed.

"Using data link, automated weather information will be available to pilots flying above 12,500 feet mean sea level and at certain airports. Request-reply weather service will be available as will significant meteorological reports from center meteorologists.

"VOR radio navigation stations will provide the routine weather broadcasts. Computer-generated voice weather broadcasts will be presented continuously over a national VOR network with coverage at 2,000 feet altitude and higher above the terrain. These automated weather broadcasts will give current weather and significant meteorological reports to the large number of general aviation and helicopter pilots who fly at the lower altitudes. Automated weather sensor voice outputs will be broadcast over VOR stations giving airport surface weather information and allowing lower weather minimums for landings. Flight service station voice communications with pilots will take place over other local and remote communication outlets. Additional automated weather sensors will be installed at airports.

"The center weather processor at air route traffic control centers will distribute current weather radar to the flight service specialist, center meteorologist, central flow control, and tower cabs at major airports. Weather radar will be distributed to displays used at the center and by the terminal radar approach controllers.

"In addition, the center weather processor will provide automated distribution of alphanumeric and graphic weather information to operating positions in flight service stations, via the flight service data processing system (FSDPS). More accurate and timely weather radar information will be provided through improved FAA terminal radar weather channels.

"Long Term (to 2000). Data link coverage will be extended downward from 12,500 to 6,000 feet altitude above mean sea level, giving automated weather service to aircraft at lower altitudes. Improved weather radar data for operational displays and automated air traffic control functions will continue as the next generation of weather radar systems, called NEXRAD, are added.

"Ongoing development programs, such as wake vortex, wind shear, and hazardous weather detection, will continue with more accurate and timely weather data presented to pilots and agency operations personnel" (Federal Aviation Administration, 1981).

ORGANIZATION OF THIS BOOK

Part I presents a historical overview of past research on air traffic controller selection. This work has occurred almost exclusively in the United States and Chapter 2, by Thomas F. Hilton and S. B. Sells, opens the discussion with an explanation for the up-to-now predominance of the United States in this area. As a result of changing circumstances, it appears that other countries are finding it necessary to follow suit. The next three chapters, by Leland D. Brokaw (Chapter 3), William E. Collins, James O. Boone, and Allen D. VanDeventer (Chapter 4), and Evan W. Pickrel (Chapter 5), review the developments in air traffic controller research, both in historical perspective and from the vantage of the contributing organizations. The early research, up to around 1960, took place in the Air Force and by contract

between the Civil Aeronautics Administration (CAA), predecessor to the Federal
Aviation Administration (FAA) and the Air Force, with highly promising results.
This work was continued and extended by psychologists at the FAA Civil Aero-
medical Institute (CAMI), which was established in 1960, where the contri-
butions of David K. Trites, Bart B. Cobb, and James O. Boone, along with
numerous others, have been of major importance. The contributions of psy-
chologists John T. Dailey and Evan W. Pickrel, of the FAA Office of Aviation
Medicine (OAM) and their colleagues, in the design of innovative new tests
that constitute the new selection battery, and in personality screening, per-
formance measurement, and other aspects of the overall selection program,
round out the research review. The final chapter in Part I, by Evan W. Pickrel,
describes the adjustments made by the FAA after the massive dismissal of con-
trollers who refused to return to work, in August 1981, following the strike
by the Professional Air Traffic Controllers Organization (PATCO). In addi-
tion to activation of the new selection battery, some major changes were made
in the structure of the controller jobs and in the training curriculum of the
FAA Academy.

The two chapters in Part II present important information about the men
and women who comprise the Air Traffic Controller Specialist occupation and
about several important aspects of their careers, namely aging, stress, and
performance assessment. In Chapter 7, on controller characteristics, John
Dailey brings together extensive information, from a variety of sources, that
has not been published previously and that furnishes significant insights
concerning the ability profiles of successful controllers. Contrary to ex-
pectation, Dailey found that a series of detailed job analysis studies,
although highly useful for the guidance of training, were of limited value
for understanding requirements for test battery design.

In Chapter 8, Joseph A. Tucker, Jr. presents information concerning con-
troller careers and aging, and on stress, that may correct widely held mis-
conceptions created in the popular literature. Based on official employment
records, he found that controllers' careers, prior to the PATCO strike, had
been longer than other federal careers and that controllers evidenced strong
staying power, with reference both to average retirement age and length of
service. With respect to stress, his analysis is that air traffic controllers
as a group score low on trait anxiety and that the controller job is not a
uniquely stressful occupation, although it makes a high demand for mental
concentration. In the final section of this chapter, Tucker discusses the
development of objective performance assessment, to enable meaningful recog-
nition and reward for good performance and to supplant the subjective methods
that have been predominant in the past.

The last theme is pursued in Part III, edited by Jack M. Greener. There
are four chapters on performance measurement during training. In Chapter 9,
James O. Boone describes the ATC training program at the FAA Academy and the
methods that have been developed for the assessment of student performance
in the Terminal and EnRoute options. This is followed by Evan Pickrel, in
Chapter 10, on assessment of student performance in the Flight Service Station
option. These discussions place emphasis on the assessment of pass-fail in
the respective courses. Chapter 11, by Evan Pickrel and Jack Greener, des-
cribes one of the objective measures developed for training assessment, the
Controller Skills Test, and Chapter 12, by Joseph Tucker, describes the devel-
opment of paper-and-pencil job simulation tests, which will eventually serve
an important function in the objective assessment of training performance as

22.

well as on-the-job proficiency. In Chapter 13, Jack Greener reviews measures that have been used to assess post-training, on-the-job performance and the prospects for system-based, computer-scored measures. His review of the state-of-the-art of controller performance assessment concludes Part III.

Edited summaries of technical reports related to the development and validation of new tests and to the development, validation, and utility assessment of the new selection battery, comprise the text of Part IV. The new tests are the Multiplex Controller Aptitude Test (MCAT), in Chapter 15, and the Occupational Knowledge Test (OKT), in Chapter 16, both by John Dailey and Evan Pickrel. Chapter 17, by John J. Convey, reports research on personality characteristics of controller applicants and the development of a psychiatric screening test that is currently in use by the FAA. The next five chapters (18 through 22) summarize studies that have for the most part appeared only in technical reports, that provide the main documentation for the new selection battery, and the final chapter -- 23, summarizes conclusions and recommendations for operational use of the test battery. With the exception of two contract studies by Joseph G. Colmen and his associates at Education and Public Affairs, Inc. (EPA), a private firm, (reported in Chapter 18), the authors of the remaining studies are Donald B. Rock, John T. Dailey, James O. Boone, and Evan W. Pickrel of FAA and Herbert Ozur of the Office of Personnel Management (OPM), formerly the Civil Service Commission. Chapters 18 through 23 were prepared in final form by S. B. Sells.

The final two chapters, in Part V, address the expected changes in the NAS, as announced in the new Plan, the expected changes in the job of the controller, assuming that the Plan is implemented on schedule, and new research on controller selection that appears to be indicated, based on the information and assumptions presented.

Chapter 24, by Neal A. Blake, discusses the new Plan, commenting on the new hardware and software that is to be developed, the time frame in which various items and procedural changes can be expected to come "on line," barring unforeseen obstacles, and the expected effects of the new equipment and procedures on the job and role of the controller in the system.

Using this and other published and unpublished information as sources, the final Chapter (25), by S. B. Sells and Evan Pickrel, attempts to characterize the job and role of the controller in the system, both in the near term (up to 1985) and beyond, and discusses new research that appears to be indicated to anticipate changes in controller selection that may be required to keep up with system changes and system requirements.

23.

PART I.

HISTORICAL OVERVIEW OF CONTROLLER SELECTION RESEARCH

Research on the selection of air traffic controllers in the United States can be described in terms of two periods, an early period related to the development of the Civil Service Commission Test Battery which was adopted for operational use in 1964 and continued in use until 1981, and later, the development of the new OPM-ATC Selection Battery, which extended from around 1970 to 1981, when the new battery was formally adopted.

The 5 chapters in Part I. discuss this research in historical perspective. Chapter 2, by Thomas F. Hilton and S. B. Sells, presents a rationale for understanding the unilateral development of scientific, actuarial methods of controller selection as a uniquely American need as well as a uniquely American solution. As a result of recent social, economic, and technologic changes, the American methods have been borrowed by other countries, and current practices in a number of countries are described.

Early research on controller selection, mainly in the military, spanning the period from 1941 to 1963, is described in Chapter 3 by Leland D. Brokaw, who had major responsibility in this area at the Air Force Personnel and Training Research Center. This period overlapped the first three years of the important work of the group at the FAA Civil Aeromedical Institute (CAMI), established in 1960, during which cooperative research by Brokaw and David K. Trites of CAMI took place, and set the stage for later work by Trites, Bart B. Cobb, James O. Boone, and other CAMI scientists that contributed significantly to both the 1964 and the 1981 batteries. The CAMI contributions were extensive and are summarized in some detail in Chapter 4, by William E. Collins, James O. Boone, and Allan D. VanDeventer.

Another organization within the FAA, that took primary leadership in the conceptualization and basic new test development for the 1981 battery, is the headquarters Office of Aviation Medicine (OAM), in which John T. Dailey and Evan W. Pickrel were principally responsible for the controller selection activities. Chapter 5, by Evan W. Pickrel, describes the contributions of the OAM group.

The final chapter of Part I, also by Pickrel, discusses the adjustments that were made in the Air Traffic Service as a result of the strike, on August 3, 1981, by the Professional Air Traffic Controllers Organization (PATCO), when over 11,000 controllers who failed to return to work were dismissed. This chapter describes several significant changes, including a major restructuring of the controller jobs, that occurred too recently to be reflected in the extensive research reported throughout this book. It also reports the most recent results of the use of the new selection battery, first administered in the fall of 1981, which were highly confirmatory of the performance expected on the basis of the developmental research.

S. B. Sells

25.

Chapter 2

AIR TRAFFIC CONTROLLER SELECTION IN THE UNITED STATES
AND OTHER COUNTRIES. AN INTERNATIONAL OVERVIEW

Thomas F. Hilton and S. B. Sells

Although the selection of air traffic controllers is a worldwide problem,
it has received little research attention outside of the United States. With
the exception of the United States, there is practically no published research
validating the standards, instruments, and procedures used to screen appli-
cants for employment or training as ATC specialists. What is currently known
about the value of certain testing instruments and background standards has
been provided by research within and sponsored by the United States Federal
Aviation Administration (FAA).

This chapter addresses some of the reasons for the paucity of validation
research on ATCS selection in other countries, and reviews the standards em-
ployed by many major European and Eastern countries in the free world.

FACTORS LEADING TO UNITED STATES DOMINANCE
IN ATCS SELECTION RESEARCH

The high degree of American involvement in ATCS selection research and
development appears to be the result of several related factors. First, the
United States became the early leader in the expansion of commercial and civil
aviation, thereby causing an early need for some form of air traffic control
on a large scale. At the same time, the United States was the dominant country
in the development of technological innovations employed in ATC operational
support. And finally, the United States has been the early leader and prin-
cipal center of psychological measurement, including pre-employment tests and
other screening instruments.

The United States Role in Civil and Commercial Aviation

The requirement for air traffic control was a result of the post World
War II boom in American civil aviation. A number of factors contributed to
this boom, including (a) a large number of experienced ex-military pilots,
(b) the availability of a large surplus of decommissioned (and affordable)
postwar aircraft, (c) a number of highly productive aircraft manufacturers,
faced with a dwindling market for war planes, and (d) a national economic
situation free of the burden of postwar reconstruction (Solberg, 1979;
Whitnah, 1966).

The large number of postwar pilots, coupled with the availability of
affordable aircraft, led to rapid growth in both commercial and civil avi-
ation traffic. Many pilots obtained loans to purchase planes for cargo and
passenger transport. Others purchased or rented planes for pleasure or per-
sonal transportation. Aircraft manufacturers also began to exploit new
commercial and civil aviation markets in hopes of maintaining their wartime
gains in productivity.

26.

By the mid-1950's, most major airports had implemented radio controlled airspace to coordinate takeoffs and landings due to increased traffic density (Gilbert, 1973). Nevertheless, there remained serious limitations to the dependability of passenger and freight service, caused by frequent deterioration of weather and the difficulty of safe nighttime air navigation. As a consequence of economic pressures for expanded commercial air service, the U.S. Civil Aviation Administration began, around the end of the 1950's, to develop improvements in its air traffic system on a nationwide scale (Frederick, 1961; Solberg, 1979).

The United States' Role in Air Traffic Control Technology

The radio traffic control system employed in the coordination of airport landings and takeoffs had proven that enroute daytime control under visual flight rules was workable. The assistance of radar approach control developed by the military during World War II had also provided a useful tool for the control of airport traffic congestion. By establishing air routes, highways in the sky, it became possible to control aircraft enroute between airports, even at night, using a system of radio beacons. Although collision-avoidance was still primarily a matter of pilot vigilance, the airspace was made sufficiently predictable to enable nighttime air traffic. This was a boon to the profitability of commercial aviation.

Had American science and industry not been ready to provide superb radio and radar technology, enroute traffic control, dependable enough to enjoy almost unlimited consumer confidence, would not have been the almost "instant success" that it appears to have been. However, the U.S. Civil Aviation Administration (CAA), predecessor to the FAA, was ready with facts and figures and proclaimed air travel as safer and faster than its most serious competitor -- trains. By the late 1950's, the CAA's campaign had critically damaged passenger rail service in favor of flying (Frederick, 1961). This resulted in continued mammoth commercial aviation expansion and continued air traffic density.

Despite gains in enroute traffic control systems, weather continued to present a serious problem in the 1950's and cancellations, delays, or rerouting were not uncommon. However, with the approach of the 1960's, United States radar technology was producing instrument landing systems capable of reducing the weather minimums that restricted timely flight arrivals and departures. This advance caused even broader use of air travel and enabled expanded radar control and bad weather flying (Whitnah, 1966). By the mid-1960's, air traffic density throughout the visual and instrument airspace was approaching an almost unmanageable level. Mistakes were increasing in frequency, and system reliability was under question (Boyle, 1975).

Again, American technology provided an answer, this time through the application of commercially developed data processing computer systems. The integration of computers within the air traffic system permitted air traffic control specialists to handle more flights with greater precision, and presumably, safety. However, this did not afford much relief to the system. Right on the heels of the installation of computerized enroute control systems in the mid-1960's, commercial jet aircraft came on line and jet air traffic began another rapid expansion.

27.

The jet aircraft placed a new burden on controllers because it could travel both higher and faster than propeller craft. This necessitated a revision of flight rules throughout the world (Burkhardt, 1967). Enroute control had little trouble in the maintenance of jet-prop separation, since the faster jets also flew at higher altitudes than their slower prop cousins. However, in the congested airspace around major air terminals, separation posed serious problems. To meet this challenge, new systems were developed that permitted interface between computers, radar, and inflight radio transmitters. These systems provided information to controllers directly on their radarscopes and thereby enhanced their ability to identify and monitor planes of various types in congested airspace (Gilbert, 1973).

American leadership in the design and manufacture of hardware for the nationwide air traffic control system ensured the expansion of our national commercial aviation industry. It also enabled the United States to advance commercial and civil aviation at a pace unprecedented in the industrialized world. However, this advancement was not entirely an engineering triumph; sophisticated technological hardware has required human participation in system operation and maintenance as well. Between 1945 and 1965 the nature of the ATCS task changed dramatically, and the system changes that occurred required associated changes in controller qualifications and training. In 1945, controllers relied on voice radio and visual contact to manage a moderate number of similar types of aircraft during daylight landings and takeoffs in clear weather. By 1965 controllers were managing large numbers of aircraft of diverse design and capabilities under positive control, 24 hours a day, in all weather conditions, using very sophisticated equipment. Rapid expansion and technical complexity posed two serious problems: (a) finding a large number of personnel with aptitudes sufficient to the mastery of the new technological systems and (b) developing methods for training those selected.

Personnel Selection and Training in the United States

Because of its geographic size and burgeoning aviation industry, the United States has experienced problems unique to its own economic success. As commercial aviation expanded, the demand for trained ATCS's began to outstrip the supply of experienced applicants, and formal selection and training programs were urgently needed.

In the early years, air traffic controllers were drawn from the ranks of personnel with military air traffic control experience. However, by the 1960's, demand seriously exceeded the available pool of experienced applicants (Cobb, Lay & Bourdet, 1971). Already in the mid-1950's research had been initiated to examine the traits of qualified ATCS's, in anticipation of increased demands for personnel. In addition to requirements to fill new ATCS slots, positions were being vacated due to retirement and the inability of some controllers to adjust to new task demands. Large scale standardized programs were established to insure minimal technical competence (Henry, Kamrass, Orlansky, and others, 1975). A centralized school and standardized aptitude screening have been in place since the early 1960's.

As the pool of experienced ATCS applicants became exhausted and the number of training failures grew to acute proportions, experience variables

and education minimums were inadequate to screen successful trainee applicants. In addition, job turnover became problematic as new hardware systems outpaced the ability of a number of older controllers to handle the pressure of operating them. It was clear that a high school diploma and interest in aviation were inadequate applicant qualities (Corson, 1970; Cobb, 1971).

Here, too, American science and technology provided means to address the developing problem. During World War II, American psychologists had demonstrated the effectiveness of methods used in screening applicants for a wide range of technological jobs, using standardized and validated multiple aptitude paper and pencil tests (Flanagan, 1947; Anastasi, 1968). These aptitude tests were inexpensive to administer, and within the limits of standardization, could be given in a variety of settings. The postwar boom in psychological testing permitted the civil aviation authorities to capitalize on this new technology to aid in solving their staffing problems.

New programs were established, based on this psychological testing technology, which enabled the selection of qualified trainees from large pools of applicants with diverse backgrounds, in order to meet staffing needs. These programs required the investigation and identification of critical variables that could predict successful training outcomes and on-the-job success for the purpose of demonstrating their validity (Brokaw, 1959; Trites, 1961). The resulting research, to which this book is dedicated, not only aided in the improvement of test utility, but it also increased knowledge of the human attributes essential to job performance.

In retrospect, the most valuable information affecting personnel decisions was an outgrowth of the research related to the development of selection variables and instruments. Maximum age limits for entry and retirement were identified, to insure job efficiency (Cobb, Young, & Rizzuti, 1976). The role of previous aviation-related experience was clarified, as were other types of experience, such as education and military service (Trites & Cobb, 1963). The kinds of cognitive and psychomotor skills essential to task mastery were explored (Cobb, 1962; Cobb, et al., 1971), and personality variables were examined (Karson & O'Dell, 1970, 1974). Many of these variables were shown to mediate training and job success, while others were dismissed as not relevant. Therefore, test research created information that went beyond the assessment of selection instruments. It also helped to understand the factors related to successful ATCS work (Cobb & Nelson, 1974).

ATCS Personnel Selection Outside the United States

As mentioned earlier, ATCS selection research and development have been virtually nonexistent in the world scientific literature outside the United States. In our opinion, this reflects three major factors: (a) slower postwar economic growth outside the United States, (b) the smaller geographic scale and complexity of most other commercial aviation systems, and (c) differences between the characteristics of ATCS applicants in the United States and other countries.

After World War II, most countries of the world suffered severe economic hardship. Only North and South America were spared destructive land battles, and only North America could claim a vital industrial economy. This delayed

29.

the rapid boom in commercial air transport elsewhere that had been experienced in the United States. Most countries were left with few undamaged aircraft manufacturing plants. Even though ex-military pilots were abundant, aircraft production was quite limited, and so was passenger demand. In Europe, especially, the small geographic size and physical proximity of the countries gave rail transportation an economic edge (Sealy, 1957). Trains were reliable, fast, and inexpensive, and they could be restored after the war without the high technology skills of aircraft mechanics.

Air transportation has always been expensive, more expensive than any land or sea transport system. Only two elements make air travel worth the cost; limited time and limited accessibility (Gronau, 1970). The old adage that "time is money" reflected the desire of salesmen and executives, operating in world markets, to hasten the pace of their schedules to meet or beat their competitors in the escalating postwar economy. For example, Danes could compete with Germans for business in Egypt, if they both could travel there in hours. Therefore, commercial aviation among industrial countries tended to develop along the lines of passenger service.

Developing countries, on the other hand, have no need for rapid passenger travel. For them, accessibility problems such as impassible terrain or large scale geography can make it cost-effective or even necessary to use air cargo rather than surface carriers. In developing countries, land-based transportation is often limited. Perishable commodities, such as medicine and foods, can best be transported over larger distances to and from inaccessible locations by air. Geological exploitation exemplifies how it is, in the long run, economical to be able to fly in small survey teams and technical equipment to determine whether resources warrant the construction of roads and rail lines.

A consequence of the interaction of economic factors and geography has been to emphasize passenger air travel in industrial nations, and air cargo in those developing countries, where passenger travel was neither in demand nor economical (Groenewege & Heitmeyer, 1964). Only in the United States did the two factors exist simultaneously -- strong industrial growth and inaccessible geographic regions (predominantly in the western United States)-- thereby encouraging both passenger and cargo air transport.

The slower, more homogeneous expansion of commercial aviation outside the United States helped to restrain its growth and to make it more manageable. As a result, air traffic control was more orderly, and technological ATC system innovations when required, had typically already been debugged, so to speak, during development in American aviation.

By 1960, commercial aviation expansion was evident in almost every country in the world. European air terminals were rapidly reaching traffic densities common to major cities in the United States, and by the mid-1960's, jet transportation insured the need for technologically advanced ATC systems throughout the industrialized world. Even in so-called "third world" countries sophisticated terminal ATC systems were necessary to insure passenger service to major urban centers in order to promote tourism and economic trade with industrialized nations.

However, the geoeconomic constraints on air transportation outside the United States held in check the staffing demands for ATC systems. Small industrialized countries had limited airspace capacity, and enroute/IFR traffic did not require large complex radar and radio nets because planes seldom exceeded the radar range of one country before being handed off to the control of another country's air traffic system (Great Britain Department of Industry, 1977). In other words, enroute flight paths were shorter and under more radar scrutiny than was possible in the United States.

Finally, apart from the smaller scale of commercial air traffic in other countries, cultural factors made selection requirements less problematic outside the United States. The United States has been somewhat unique in its political and cultural principles of equal opportunity for all citizens. In particular, the principle of universal education has provided that all citizens are entitled to education through the 12th grade. However, unlike the competitive examinations that are required in the British and European-type educational systems, where continued education is contingent on demonstrated aptitude and mastery of the prescribed curriculum, American schools have evolved a policy of "social promotion." Under its various manifestations, this policy has enabled students to obtain high school diplomas based on criteria ranging from scholarly achievement to mere persistence in attendance (Carver, 1981).

When screening is based initially on educational certification, in a country using a strict examination system, the air ministry (and other employers) can depend on the reliability of educational quality control to provide a minimal aptitude level. In the United States, on the other hand, a diploma provides considerably less differentiation between candidates. In some cases, it does not even insure literacy (Barzun, 1981). Therefore, the heavy reliance, in the rest of the world, on interviews as a method of screening has not affected selection as seriously, since most candidates certified as having the equivalent of a secondary school education could be assumed already to possess considerable aptitude based on educational achievement.

Cultural homogeneity in most countries outside the United States has also contributed to increased training effectiveness, in that failures in training due to cultural differences between the trainees and those who develop training curricula are likely to be fewer. In the United States, egalitarian policies require special consideration of persons of both low socioeconomic status and ethnic minority status. In addition, cultural and class differences between candidates and personnel officers (outside the United States) have been shown to result in less favorable evaluation of applicants (McGuigan, 1979). This is believed to have reduced the likelihood of selecting trainees of diverse ethnic backgrounds, thereby reducing the likelihood of failure due to a confusing curriculum.

Social and economic differences between the United States and many other countries also affect relative job turnover. In most countries, only well educated people have high job mobility. However, for many years, the United States has offered many more job opportunities than other countries, and mobility has been comparatively high at all levels; this has applied to air traffic control trainees and controllers as well as to other occupations.

31.

In view of the arguments outlined above, it appears that ATCS recruiting, selection, and training problems in the United States reflect different underlying factors than those in most other countries. These involve a greater dependence on testing and screening devices than appears to be necessary in other countries, and also greater emphasis on retention of personnel, in view of the mobility of the working population.

One country with a pattern of civil aviation expansion similar to that in the United States, and with similar cultural diversity, geographic scale, education system, and economic history, is Canada. Until the late 1960's Canada's commercial aviation expansion was slower but parallel to that of the United States (Manchester, 1968). It was slightly slower in development, possibly as a result of having a smaller postwar population, which was spread thinly across a large geographic area. However, with a rising birth rate and an influx of immigrants, the 1960's saw Canada come into its own as an economic power in the world. Through its close ties with its southern neighbor, the United States, expansion of the air traffic system was timely and nearly "state of the art" by the early 1970's. Air traffic density in Canada was also beginning to parallel that of the United States, with a large number of jet aircraft, mixed with private and commercial propeller craft.

About this same time, Canada began to realize trainee attrition as experienced earlier by the United States. By the late 1970's Canada began to look at its selection procedures seriously in the hope of meeting its problems through the utilization of standardized, validated screening instruments developed and employed by the United States (McGuigan, 1979).

The similarity of experience of the United States and Canada support the prediction that similar problems may develop in other countries as they continue to expand their commercial aviation and as population diversity and related problems become more acute (Harrison, 1975). Recently, the International Federation of Air Traffic Controllers Associations (IFATCA) established a review of worldwide aptitude testing procedures as the sole agenda of its Standing Committee on Training (SCT) for 1982 (IFATCA, 1981).

PROGRESS IN WORLDWIDE ATCS SELECTION METHODOLOGY

A major consequence of the factors unique to America has been an ongoing commitment to research and development in personnel selection of ATCS's for nearly two decades. This research did not go unnoticed by the air ministries of other countries, despite their lack of activity in selection aptitude testing.

Countries with relatively homogeneous populations and smaller manpower requirements have apparently not found it necessary to develop sophisticated selection tests. The impression has been gained that they consider their procedures to have functioned acceptably. Nevertheless, research in the United States has had some impact on the screening procedures in a number of countries.

Foremost has been the finding that as controllers advance in age, their performance becomes less effective (Cobb et al., 1976). As a result of this finding, and the then held belief that ATCS work is more psychologically and physiologically stressful than other occupations, the Congress of the United States passed Public Law 92-297 in 1972. This law set the normal retirement

32.

age for active ATCS control at 55 (and even earlier, with 18 years on the job). To control aircraft beyond this age required special permission based on excellent performance, and in no case beyond age 60. Table 1 presents data on the maximum age of ATCS's for a number of countries for which information was available. Fully half of these retire operational controllers before the widely accepted age of 65.

Within the trainee population, older trainees have been shown to have higher attrition rates both during and after training (Cobb, Nelson, & Mathews, 1973) and lower job performance ratings than their younger classmates (Cobb & Nelson, 1974). Hopkin (1979) has suggested that older controllers seem to be less flexible in learning ATCS work and in adapting to new systems and procedures resulting from technological advances. Therefore, in light of the shorter career length for active air traffic control mandated by early retirement, and as a result of the findings cited above, the FAA established regulations in 1973 limiting the maximum ATCS entry age to 30 years. Subsequently, other countries have begun to limit entry age, typically ranging between the mid-twenties and mid-thirties.

The role of prior experience and education has also been investigated in the United States, and with the exception of radar-IFR controller experience, other experience variables and education have been found to be unreliable indicators of ATCS training and job success. The requirement for experience as aircraft pilot has often been demanded by the pilots being controlled; this and other aspects of aviation and controller experience have nevertheless shown little relationship to competence in job performance (Cobb & Nelson, 1974). In addition, education above high school graduation has proven unrelated (or sometimes negatively related) to trainee performance. It may be that a college degree may lead some trainees to perceive ATC work as insufficiently challenging or meaningful (Cobb et al., 1976). College training also provides such individuals with increased job mobility. Despite this, because ATC work is continuing to become more technologically sophisticated, some countries are beginning to push for 14 or more years of education in anticipation of a more rigorous curriculum and the development of yet more complex automated ATC systems.

CURRENT SELECTION PRACTICES IN OTHER COUNTRIES

Based predominantly on research in the United States and the experience of the respective governments, the standards for ATCS's are today relatively consistent for all nations. Table 1, in the preceding section, reflects this with respect to retirement age.

At present, only the United States employs a comprehensive aptitude battery validated on large samples in training and on the job, to screen job applicants. Such information as we have been able to obtain indicates that the current FAA-OPM selection battery stands alone in having up-to-date published validity research to recommend its use. Canada is working toward development of a selection battery, but at present is using a version of the United States Multiplex Controller Aptitude Test (MCAT) and the Occupational Knowledge Test (OKT), on an experimental basis (Note, 1).

33.

Table 1.

Age Limits for Active Controllers in Selected
Major World Countries*

Country	Maximum Age
Austria	65
Belgium	60
Canada	65
East Africa	55
Finland	65
France	55
West Germany	53
Greece	65
Hungary	55
Ireland	65
Israel	65
Norway	65
Portugal	70
South Africa	65
Spain	65
Sweden	63
United Kingdom	60
United States	55
Venezuela	60
Yugoslavia	60

*Based on data published by The International Federation of Air Traffic
Controllers Associations (IFATCA) in the second edition of its Information
Handbook (1979), and by informal surveys conducted by The Institute of
Behavioral Research (IBR) in 1979 and 1982.

Recent information obtained from six major countries indicated that the pre-employment interview is relied on heavily for personnel screening. Nevertheless, some testing is being conducted or considered. Table 2 summarizes the results of a limited inquiry. These data reflect that one or more general aptitude tests are frequently used. In most cases, however, these are tests of limited use and unpublished validity for use in screening controllers. It is questionable that such tests are very useful for discriminating ATCS candidates, and few countries rely heavily on them. Based on FAA research, the cognitive capacity to interpret radar scope information seems to demonstrate the best discriminatory power to date (Dailey & Pickrel, 1977; Boone, 1979a). The MCAT was developed for this purpose, and some form of ATCS simulation is common during the interview phase of several countries. The United Kingdom, Canada, and the Netherlands are each currently experimenting with MCAT-type test protocols.

Table 2 also suggests that a maximum entry age is used, while prior experience is receiving limited weight, which is consistent with the American research reported above.

Finally, the United States has pushed for increased minority representation in an otherwise racial majority, male-dominated occupation, but with limited success, especially in the case of racial minorities. However, the example of the United States seems to be reflected in increased willingness of most countries today to accept female candidates into the ATCS ranks.

SUMMARY AND CONCLUSIONS

ATCS selection research and development has been slow in developing outside the United States, apparently reflecting a number of historical, economic, social, and geographic factors. Nevertheless, the continued expansion of ATC systems and increased job complexity throughout the world are beginning to have impact as the need to consider the use of standardized selection procedures and criteria is becoming recognized. Social pressures from controller organizations are also building pressures for installation of personnel screening procedures of demonstrated validity and fairness to applicants. This will probably have favorable influence in view of the high content and criterion-related validity of recently developed tests, such as the Multiplex Controller Aptitude Test and the Occupational Knowledge Test.

An informal survey of major European and Eastern countries conducted by IBR in 1979 reflected widespread interest in undertaking selection research. A study conducted by IFATCA in 1979 reported that most countries use some form of general intelligence or aptitude testing, although the majority rely heavily on nonspecified interview criteria. An inquiry similar to that of IBR by Canada's Air Transport Ministry (McGuigan, 1979) supported the IBR and IFATCA conclusions. In each survey, validation studies of aptitude instruments seem to be almost nonexistent.

The International Civil Aviation Organization (ICAO) is a likely organization to coordinate standardized applicant screening, since it already standardizes ATCS qualification and licensing internationally. In the long run, this may be a prudent step. It could also enable centralized ATCS training. This could make licensing practices more valid and effective, as well as economical.

35.

Table 2

Current ATCS Selection Practices in Selected Countries

Country	Min. Entry Age	Max. Entry Age	Sex	Years Education	Tests Aptitude	Tests Intelligence	Interviews	Prior Experience
Australia	21	36	M&F	12	Yes	Yes	Motivation, verbal expression, general suitability	Prefer pilot, air radio, ground control
Canada	18	35	M&F	12	Yes	No	Verbal expression, motivation, responses to simulated ATC problems	Prefer prior ATC assistant, radio operator
Netherlands	21	26	M&F	12-14	Yes	Yes	Psychological: stress tolerance, occupational knowledge	None required
Sweden	19	c.27	M&F	12-14	Yes	No	Stress tolerance, motivation, cooperation, initiative, responses to simulated ATC problems	Prefer pilot or air navigation
Switzerland	18	25	M&F	12-14	Yes	No	3-dimensional perception, logic, responses to simulated ATC problems	ATC assistant apprenticeship, prefer
United Kingdom	18	34	M&F	12	Yes	No	Logic, motivation, stress tolerance	None required
United States	20	30	M&F	12-16	Yes	No	Logic, motivation, stress tolerance	12 years education, plus 3 years management experience or prior ATCS or aviation-related work

36.

Finally, the most likely reason for standardized selection research and development will originate from political forces arguing for the increased appearance of fairness and objectivity in screening of ATCS applicants. Most ICAO member countries also have affiliate organizations in IFATCA, which have become increasingly vocal and are not without influence in many of these countries. Personnel selection methods may well become a more sensitive issue in coming years; and this is an area where most countries are vulnerable to the criticism of benign neglect. The United States has laid the groundwork for such research and it seems that the world community could readily exploit this opportunity.

REFERENCE NOTE

1. Derived from phone conversation between Evan Pickrel and Dennis Kirby, Canada Air Ministry, December 1981.

Chapter 3

EARLY RESEARCH ON CONTROLLER SELECTION

1941 - 1963

LELAND D. BROKAW

The selection of air traffic control personnel has been a matter of con-
tinuing study since 1941. Constantly increasing air traffic, technological
change, and an imposed legal requirement for demonstrable job relevance have
contributed impetus to the research. Mahlon V. Taylor (1952), in a report of
research by the American Institute for Research (AIR), cited an unpublished
study by Dewey Anderson in 1941 in which 28 tests, including 17 Thurstone
Primary Mental Ability measures, were given to most controllers in the mili-
tary service. With job performance ratings as criteria, the four highest
validities were attained by a measure of perceptual speed and accuracy en-
titled "Three Higher," a space relations test called "Flags," and tests of
reasoning and integration entitled "Letter Series," and "Pedigrees." These
factor areas were included in a test battery assembled by AIR in 1951 in the
accomplishment of a contract for the Civil Aeronautics Administration (CAA),
to develop and validate tests to screen applicants for jobs in air traffic
control (Note 1).

The AIR Contract Study

In addition to the measures designed to probe the factorial areas found
valid in the earlier work, predictor tests were prepared to address specific
content areas identified in a job analysis, and others were designed to serve
as job samples. AIR personnel solicited critical incidents from supervisors
in both air route control centers and control towers. The critical activities
and accompanying aptitude components identified by AIR appear in Table 1.
Because of a Civil Service Commission policy that restricted selection tests
to paper-and-pencil measures, measures of auditory perception, writing rapidly
and legibly, copying behind, speaking intelligibly at optimal speed, verbal
fluency, and the long term aspects of memory could not be included.

In all, 18 tests were included, that measured the aptitude components
as they might occur in real job situations. After preliminary tryout and
refinement, the test battery was administered to 211 persons, 90 from Air
Traffic Control Centers, 75 from Terminal control towers, and 46 from communi-
cations stations. Applied research in operational environments often encounters
difficulty in obtaining samples of subjects for study and this work was no
exception. The 90 persons from centers were tested in five locations; the 75
from towers, in 12 locations; and the 46 from communications stations, in
seven places. These small samples made assembly of comparable criterion data
difficult; indeed, the accompanying loss of cases made it impossible to vali-
date within the communications samples. Nevertheless, a process of validating
within each subsample and deriving average validity correlations through
Fisher's z-transformation provided analytic samples of useful size.

Table 1

Critical Activities and Incidents and Aptitude Components for Air Route Traffic Control (R), Airport Traffic Control (P), and Aircraft Communicator (Air-Ground) (C)[a]

Critical Activity and Aptitude Components	Percent of Critical Incidents[b]		
	R	P	C
1. Receiving Oral Messages: Auditory Perception Verbal Comprehension	8	13	22
2. Recording Oral Messages: Writing Rapidly and Legibly, Copying Behind, Encoding, Memory for Interrupted Tasks	9	9	23
3. Recording Self-Originated Data: Visual Perception, Carefulness	9	6	9
4. Displaying Flight Data: Memory for Interrupted Tasks, Visual Perception, Carefulness	11	8	1
5. Requesting Information:	8	11	12
6. Coordinating (Clearances, etc.): Memory for Interrupted Tasks, Carefulness, Integration, Assimilating Symbolized Data, Comprehending Unseen Movements	21	11	3
7. Devising Clearances: Carefulness, Integration, Short Term Memory, Judgment, Visual Perception of Time-Distance Relationships, Assimilating Symbolized Data, Comprehending Unseen Movement	11	8	0
8. Devising Taxiing, Takeoff, and Landing Instructions: Carefulness, Integration, Short Term Memory, Judgment, Visual Perception of Time-Distance Relationships, Assimilating Symbolized Data, Comprehending Unseen Movement	2	13	0
9. Issuing Oral Communications: Carefulness, Verbal Fluency, Speaking Intelligibly at Optimal Speed	10	10	13
10. Evaluating Priority of Communications: Judgment, Integrations, Short Term Memory	7	11	17

a. Taken from Taylor 1952

b. These are percents of all critical incidents reported for a given job during the test validation (799 for Air Route Traffic Control, 1193 for Airport Traffic Control, 150 for Communications)

40.

Several kinds of data were collected in conjunction with development of the criterion measures. Biographical data and existing official evaluations were obtained for personnel of the facilities at which tests were given. In addition, supervisors were requested to record incidents of critical performance on the Airways Operations Performance Records, which has been developed for the job analysis, and to provide overall performance records of personnel who were tested.

Quoting from Taylor (1952, pp. 11-13) "In general all the measures were useful as criteria, although there was considerable variability among facilities in this respect, and it was felt advisable to select and combine, for each of the seven facilities, the measures (aside from supervisors' ratings) proving most satisfactory for the given facility. The correlations between supervisors' ratings and these composite measures averages .81, indicating that both....were measuring the same thing."

As indicated above, when the tests were administered in the field, practical difficulties made themselves felt. Both personnel to be tested and testing time were in short supply; as a result, the sample sizes were adequate for determination of validities for only fourteen of the tests. Table 2 presents the validities obtained for sixteen of the tests, but tests numbered 12 and 13, because of very small sample sizes, were not considered when the recommended battery of nine tests was selected.

In summarizing the results of the study, Taylor (1952) indicated that although the battery was suitable for use with candidates who had prior experience in flying, or in some controller function, he felt that use of the tests with naive candidates would require extensive revision of the directions by experienced test development personnel, and that additional experimental testing would be required.

Although the AIR delivered a recommended battery of tests for the selection of air traffic controller trainees, it was not implemented.

The next study resulted from a visit by CAA representatives to the Personnel Laboratory, Air Force Personnel and Training Research Center, at Lackland Air Force Base, Texas, in February 1956. Their visit was prompted by a desire to revise selection procedures for trainees in the CAA Air Traffic Control School. Up to the time of their visit, air traffic controller selection was based upon previous on-the-job experience and a physical examination. A foreseen need to increase the numbers of trainees imposed the requirement for a method of selection from a naive population.

Joint Air Force/Civil Aeronautics Administration Study.

The visit of the CAA representatives to the AF Personnel Laboratory resulted in a joint Air Force/Civil Aeronautics Administration study. A battery of 37 tests, yielding 90 scores, was administered to entering trainees in the ATC school (Brokaw, 1957). The battery was heterogeneous, selected to cover most of the variance believed relevant to success in training. Commercial tests, Air Force tests, and the more effective tests devised by the AIR were included. Initial validation was accomplished against available training criteria, including an average lecture grade, individual instructor ratings

41.

Table 2

**Air Route Traffic Control and Airport Traffic Control Average
Test Validities with Composite Criteria and with
Supervisors' Ratings**

Test	Composite Criteria		Supervisors' Ratings	
	r	Nb	r	N
1. Locating Data I	.06	71	.09	83
2. Locating Data II	.06	71	.08	71
3. Air Traffic Math I	.05	71	.06	83
4. Air Traffic Math II	.04	54	.19	54
*5. Memory Flight Information	.20	53	.24	53
*6. Air Traffic Problems I	.49	52	.51	64
*7. Air Traffic Problems II	.53	52	.43	64
*8. Flight Location	.18	62	.32	74
*10. Coding Flight Data I	.29	68	.29	80
12. Taxiing Aircraft	.21	19	.28	19
13. Control Judgment	.39	37	.36	37
*14. Memory for Aircraft Position	.33	76	.22	88
15. Three Dimensional Visualization	.22	27	.08	41
*16. Circling Aircraft	.38	44	.28	56
*17. Aircraft Position	.28	44	.36	56
*18. Flight Paths	.17	44	.37	56

a. Data extracted from Taylor 1952
b. The rs are as reliable as if computed from a single sample with the given N.
* Tests included in the recommended battery.

42.

of student performance, and a composite instructor rating based upon dis-
cussions among the three or four instructors who had dealt with each class.
There were about 20 students in each class; of 197 students who completed
the experimental battery, training criteria were accumulated for 130.

In addition to tests covering aptitude factors believed relevant, and
some job-sample tests, there were measures of temperamental factors seem-
ingly relevant to the ATC task. Temperament tests included were the Cali-
fornia Test Bureau (CTB) Occupational Interest Inventory, CTB Mental Health
Analysis, CTB Test of Mental Maturity, and the CTB California Test of Per-
sonality. Aptitude tests included were the CTB Survey of Space Relations
Ability, CTB Survey of Working Speed and Accuracy, CTB California Capacity
Questionnaire, CTB Personnel Selection and Classification Tests, the Differ-
ential Aptitude Tests published by the Psychological Corporation, and United
States Air Force tests of numerical, verbal, reasoning, mechanical, and
spatial factors. Job sample tests were the AIR Air Traffic Problems and
Locating Data.

This study was intended to provide a test battery to replace a previous
system involving experience variables. In that context the contribution of
experience to school success was also sought. Background and experience
variables included in the study were age, educational level, marital status,
and previous air traffic experience. Experience was studied in the following
five categories:

A. Any air traffic control experience versus no such experience.
B. Experience in air traffic control versus no such experience.
C. Experience in ground control approach versus no such experience.
D. Senior CAA ratings versus no such ratings.
E. CAA certification versus no such certification.

Validation correlations were computed against the three training cri-
teria (1. Average lecture grade, 2. Average instructor rating, 3. Composite
instructor rating). All of the variables included in the validation study
are presented in Appendix A. None of the temperament measures showed useful
levels of prediction and these results have been omitted from the tables.
The validities for the aptitude and job sample measures are reported in Table 3.
Although the measures were not uniformly significant in prediction of the
three criteria, all measures showed useful levels of validity for at least
one of the criteria, and most measures were valid at a significant level for
all of the criteria.

The background and experience variables were moderately related to the
three criteria, although it was found that educational level and previous
flying experience were not predictive. Age was negatively related at a sig-
nificant level, while ATC experience variables were generally relevant.
The validities of the background and experience variables are presented in
Table 4.

Multiple correlations were computed to evaluate the contribution of the
various measures to the prediction of school success. Two criteria were
chosen for this purpose: the average lecture grade, and the composite instruc-
tor rating. These data appear in Table 5. It should be noted that only the

43.

Table 3

Validities of Experimental Tests for Three
Air Traffic School Criteria[a]

N = 130

Content Area Test	Validity[b] 1[c]	2	3
Computational and Arithmetic Reasoning			
Dial and Table Reading (Air Force)	.43	.38	.38
California Capacity Questionnaire (6)	.17	.22	.24
Number Series (Mental Maturity)	.21	.18	.23
Numerical Quantity (Mental Maturity)	.28	.21	.24
Arithmetic, Personnel Selection & Classif.	.33	.28	.32
Numerical Ability, DAT	.32	.28	.31
Air Traffic Problems I, AIR	.30	.31	.37
Arithmetic Reasoning (Air Force)	.38	.27	.26
Perceptual and Abstract Reasoning			
Calif. Capacity Questionnaire (5)	.33	.29	.28
Abstract Reasoning, DAT	.18	.15	.20
Space Relations, DAT	.21	.15	.20
Aerial Landmarks, Air Force	.16	.27	.27
Spatial Orientation, Air Force	.13	.22	.18
Instrument Comprehension, Air Force	.26	.23	.23
Verbal Tests			
Calif. Capacity Questionnaire (7)	.18	.18	.18
Reading, Personnel Selection & Classif.	.28	.16	.16
Language Usage, Sentences, DAT	.22	.18	.15
Verbal Test, Air Force	.21	.15	.13
Perceptual Speed and Accuracy			
Code Translation, Calif Test Bureau	.27	.24	.27
Counting, Calif Test Bureau	.29	.24	.19
Temperament Tests			
Family Relations, CTB, Cal. Test of Personality	.09	.19	.15
Nervous Manifestations, CTB, Cal Test of Personality	.08	.07	.06

a. Taken from Brokaw 1957

b. Correlations of .17 are significant at the 5% level; .23 at
   the 1% level.

c. Criterion 1 is average lecture grade, 2 is independent instructor
   rating of overall performance, 3 is an instructor composite rating
   based upon the joint judgment of three or four instructors
   teaching each class.

Table 4

Validities of Selected Background and Experience Factors
for Three Air Traffic Control School Criteria[a]

N=130

Variable	Mean	SD	1[c]	Validity[b] 2	3
Age	26.21	4.20	−.04	−.24	−.24
Education	12.59	1.22	.16	−.11	−.07
Martial Status[d]	.68	.46	.16	.05	.08
Previous Flying Experience	.17	.37	.09	−.10	−.11
Airport Traffic Control	.65	.48	.22	.24	.24
Ground Control Approach	.41	.49	−.06	.16	.14
Any Air Traffic Experience	.78	.41	.04	.23	.24
Senior CAA Rating	.91	.40	.26	.26	.24
CAA Certification in Any Status	.41	.49	.28	.39	.38

a. Taken from Brokaw 1957
b. .17 significant at the 5% level, .23 significant at the 1% level.
c. Criterion 1 is average lecture grade, 2 is independent instructors'
rating, 3 is a composite rating based upon the joint agreement of
three or four instructors teaching each class.
d. Marital Status and subsequent entries are dichotomized and
computed as point-biserial correlations.

45.

Table 5

Most Efficient Combinations of 2-5 Variables for Prediction of Two Criteria[a]

N=130

Variable	Val	Beta Weights[b]							
		2-Var		3-Var		4-Var		5-Var	
Average Lecture Grade									
Arithmetic Reasoning	.43	.38	.38	.33	.30	.36	.31	.28	.30
CAA Certification Status	.32	.23		.22		.26		.23	
Air Traffic Problem I	.30		.21	.20	.22	.27	.29	.29	.26
Symbol Reasoning	.34				.21		.26	.22	.26
Locating Data AIR	-.02					-.22	-.22	-.26	-.24
Code Translation	.24								.15
Multiple Correlation[c]		.49	.48	.52	.51	.56	.55	.59	.57
Composite Instructor Rating									
Air Traffic Problems I	.36	.33	.31	.29	.31	.24	.27	.24	.26
CAA Certification Status	.37	.33		.30		.32		.29	
Arithmetic Reasoning	.32		.28	.24	.22	.20	.20	.17	.21
Symbolic Reasoning	.28				.20		.20	.15	.19
Code Translation	.27					.17	.13	.16	.12
Family Relations	.26								.14
Multiple Correlation		.49	.45	.54	.49	.56	.51	.58	.53

a. Taken from Brokaw 1957
b. In each pair of columns the first selection is from all variables, the second is from test variables only.
c. All significant at the 1% level.

general variable of CAA certification status was sufficiently unique to appear in the multiple correlations. Its zero-order validity was of the same magnitude as those for the test variables, and its contribution was significant.

In May 1957, the trainees who had received the experimental battery were followed onto the job, and various criteria of their work performance were collected (Brokaw, 1959). The ratings collected from the supervisors were parallel to those collected from the instructors during the training phases of the study. The CAA also collected data on the time spent on the job by each controller before he was recommended for promotion from the trainee level to the helper level. The correlation between the instructor rating (during training) and the supervisor rating (on the job) was .59; that between the average training school academic grade and the supervisor rating was .33. In the sample of 133 controllers these values were both significant at the one-percent level. The time elapsed before recommendation for promotion peaked around the third and fourth month on the job. There was very little variance on this measure and the associated correlation was .18 with the supervisor rating, indicating that it was of little use for this study.

The higher validities among the experimental tests were found for the arithmetic reasoning, computational, and numerical speed tests. The AIR Air Traffic Problems test showed acceptable validity for both academic criteria and supervisor ratings. The majority of tests in this group were predictive of the supervisor ratings.

The abstract reasoning and perceptual tests showed useful levels of predictive efficiency for the supervisor ratings, but the verbal and clerical speed and accuracy measures were not significantly related to this job performance criterion. The background and experience variables tended to be only marginally predictive of job performance.

A comparison of the multiple validities of four tests in relation to instructor ratings and supervisor ratings appears in Table 6. Although the validities are not spectacular, they are significant and indicate that the proposed battery would have had potential for prediction of performance on the job.

Federal Aviation Administration Follow-up of AF/CAA Study

The data base developed by Brokaw in the 1956-57 period provided predictor variables for a follow-up study by Trites, in 1961. The Trites study addressed the extent to which the data would be predictive of job performance, retention in air traffic control work, and medical history over a five-year period.

As reported by Trites (1961), "Regional offices of the Federal Aviation Agency were able to supply current FAA facility addresses, or other information on all but 10 of the original 197 subjects.....Of the remaining 187 subjects, 16 had failed the training course and left the FAA early in 1957, 15 had passed the training course and had left the FAA, two were deceased, replies were not received for two, and three were with the FAA but no other information was available. This left 149 subjects (including four training course failures still with the FAA) for whom relatively complete criterion data were obtained."

47.

Table 6

Multiple Validity of Selected Tests for Training and On-the-Job Criteria[a]

Test	Instructor Ratings (N=130 Students)		Supervisor Ratings (N=133 Controllers)	
	Beta Wt	Validity	Beta Wt	Validity
Air Traffic Problems	.27	.36	.20	.25
Arithmetic Reasoning	.20	.32	.11	.23
Symbolic Reasoning and Perceptual Speed	.20	.28	.16	.22
Code Translation	.13	.27	.05	.14
Multiple Correlations	.51[b]		.34[b]	

a. From Brokaw 1959
b. Significant at the 1% level

48.

The performance criterion measures collected included (1) average supervisory rating, (2) active vs inactive controller, (3) with the FAA vs not with the FAA, (4) mean hours of sick leave, (5) no symptoms vs symptoms, and (6) no disciplinary action vs disciplinary action. Definitions of these measures, extracted from Trites (1961), appear in Table 7.

Using individual predictor data provided by Brokaw from his 1957 and 1959 studies, Trites ran multiple regressions of psychological test and biographical variables against the academic and supervisory criteria defined and applied by Brokaw. Two equations were derived for each criterion, the first based upon the psychological test variables defined in Table 3, and the second involving these psychological variables and the biographical measures of age, education, and marital status.

Two of the equations involving all the variables were found to be identical, so a sixth equation was derived involving the psychological tests, the three biographical variables, the 1956-1957 average lecture grade, and the composite instructor rating, as predictors of the 1957 average supervisor rating (Trites, 1961).

Finally, a seventh equation was derived by approximating raw score regression weights for the predictor set recommended by Brokaw (1959). The results of the first six equations are presented in Table 8. The regression coefficients appear in Table 9.

The significant correlations between the derived composite scores and the criterion measures described in Table 7 appear in Table 10. Partial correlations were computed to eliminate the effects of age. Regression equation 2, alone, was significantly related to the disciplinary criterion, and that was at a relatively low level. Regression equations 5, 6, and 7 were significantly related to the 1961 average supervisor rating.

According to Trites (1961), "If we note the correlations between the 1956-1957 criterion measures and the 1961 criteria....(see Table 11)....it is not surprising that Regression No. 6, which included the Composite Instructor Rating as a predictor variable had the highest correlation with the 1961 Average Supervisor Rating. What was surprising was the remarkably high correlation between the Composite Instructor Rating and the 1961 Average Supervisor Rating. Obviously, the instructors in 1956 were making exceptionally valid judgments concerning a trainee's potential for air traffic control work."

It was also noted that the composite recommended by Brokaw in 1959 remained significant, leaving no doubt that a valuable contribution can be made to the selection of air traffic controllers by the use of psychological tests.

Two biographical variables, Previous Flying Experience, and Any Air Traffic Experience were significantly related to the 1961 Average Supervisor Rating, but when the effect of age was removed, these relationships became nonsignificant. Thus, none of the biographical variables representing previous experience was related to subsequent job performance. It appears that the experience variables may make a valuable contribution to selection for training, but have little impact upon performance after completion of the school and some years of experience.

49.

Table 7

Description of the 1957 and 1961 Criterion Measures

Criterion Variable	N	Mean	Std. Dev.
1957 — Average Lecture Grade: The Lecture Grade Attained by Each Student in the Training Course.	195	92.4	4.20
1957 — Composite Instructor Rating: A Numerical Transformation of the Consensus of ATC Instructors Relative to the Student's Standing in Nine Topical Areas Relating to Work Habits, Ability, Emotional Stability, and the Exercise of Judgment and Reasoning. Ratings Were Obtained After Briefing Instructors on Rating Theory, Halo Effect, and Objectivity.	188	9.93	5.42
1957 — Average Supervisor Rating: A Numerical Transformation of the Consensus of Supervisors Relative to the Individual's Standing on Eleven Areas Relating to Work Habits, Ability, and the Exercise of Judgment and Reasoning. Form Was Almost Identical to One Used by Instructors Except for Two Additional Items Concerned with Demonstrated Aptitude for ATC Work and Potential Ability. Forms and Rating Instructions Were Supplied by Mail	170	27.1	5.95
1961 — Average Supervisor Rating: A Numerical Transformation of Ratings Collected from 1 to 4 Supervisors of Each Individual on a Form Containing the Same Items Used for the 1957 Average Supervisor Rating. Three New Items Related to the Individual's Emotional Stability and Relationships to Others Were Added. Forms and Rating Instructions Were Supplied by Mail. (Using Individuals with 2 or More Forms, a Corrected split-Half Reliability of .75 Was Obtained for the Derived Scores.)	149	33.9	8.37
1961 — Active vs. Inactive Controller: Individuals Were Dichotomized as Being Either Still Active in Controller Work or Not Active.	169	.822	.382
1961 — With the FAA vs. Not with the FAA: Individuals Were Dichotomized as Being Either with the Federal Aviation Agency or Not with the FAA.	169	.882	.323
1961 — Mean Hours of Sick Leave: The Mean Number of Hours of Sick Leave Taken in the Years 1957 Through 1960 Was Computed for Each Individual.	142	41.1	34.9
1961 — No Symptoms vs. Symptoms: A Dichotomy Representing Medical Complaints of Individuals as Known to and Reported by their Facility Chiefs.	146	.829	.377
1961 — No Disciplinary Action vs. Disciplinary Action: A Dichotomy Representing the Presence of or Absence of Disciplinary Actions Taken as a Result of Control Errors. Reported by Facility Chiefs of Individuals.	148	.824	.380

Taken from Trites, 1961

Table 8

Most Efficient Combinations of Psychological Tests, Biographical Variables, and the 1957 School Criterion Measures for Prediction of Average Lecture Grade (Equations 1 & 4), Composite Instructor Rating (Equations 2 & 5), and the 1957 Average Supervisor Rating (Equations 3 & 6)

No.	Variable Description	Psych. Tests & Biog. 1 B[†]	1 r[‡]	2 B	2 r	3 B	3 r	Psych. Tests Only 4 B	4 r	5 B	5 r	Psych. Tests, Biog. & 1957 School Criteria 6 B	6 r
Test Variables													
1	Dial & Table Reading (AF)	20	40**	27	37**	22	33**	14	40**	25	37**		
2	Verbal Knowledge & Reasoning (Test 6, Cal. Capacity Quest.)					23	30**					28	30**
7	Air Traffic Problems I	17	25**	25	32**			16	25**	24	32**		
8	Arithmetic Reasoning, AC 2A	24	39**					21	39**				
9	Symbolic Reasoning & Perceptual Speed (Test 5, Cal. Capacity Quest.)	21	35**	20	27**			19	35**	19	27**		
10	Abstract Reasoning, DAT					22	27**					23	27**
12	Aerial Landmarks, AFOQT					-18	07					-17	07
16	Reading, Pers. Select. & Class. Test							16	28**				
21	Family Relations, CTB, Cal. Test of Pers.					19	22**					50	22**
22	Nervous Manifestations, CTB, Mental Health Analysis	21	22**					26	22**				
Biographical Variables													
25	Marital Status	16	09*	16	09*								
1957 Criterion Variables													
	Composite Instructor Rating											60	56**
	Multiple Correlations	59**		50**		48**		59**		47**		65**	

[1] Intercorrelation matrix for equations was based on a common N of 135 trainees who completed the training course. All decimal points have been omitted from table entries.

[†] Regression Coefficients (Betas) for standardized scores.

[‡] Validity Coefficients.

*Significant at less than the .05 level.

**Significant at less than the .01 level.

°Point biserial correlation

Taken from Trites, 1961

51.

Table 9

Regression Coefficients for Standardized Scores (Betas) Reported by Brokaw and Used to Derive Equation Number 7[*]

| Variable | | |
No.	Description	Regression Coefficients[†]
7	Air Traffic Problems I	20
8	Arithmetic Reasoning, AC 2A	11
9	Symbolic Reasoning and Perceptual Speed (Test 5, Cal. Capacity Quest.)	16
19	Code Translation, Survey of Working Speed and Accuracy	05

[*]Table entries were taken from: Brokaw, L.D. School and job validation of selection measures for air traffic control training. WADC-TN-39, Pers. Lab., WADD, USAF, Lackland AFB, Texas, 1959.

[†]Decimal points omitted.

Taken from Trites, 1961

52.

Table 10

First Order and Partial Correlations Between the 1961 Criterion Measures, the 1957 Criterion Measures, and Scores Predicted from Regression Equations Where the First Order Correlations with the 1961 Criteria Are Significant

Criterion	Variable	Criterion r¹	Criterion N	Age r	Age N	Criterion r with Age Partialed Out
1961 Average Supervisor Rating	Age	−23**	149			
	Average Lecture Grade	24**	148	−05	135	23**
	Composite Instructor Rating	45**	149	−18*	135	43**
	1957 Average Supervisor Rating	33**	143	−07	135	32**
	Regression Equation 5	17*	143	−14	133	14
	Regression Equation 6	44**	127	−09	133	43**
	Regression Equation 7	23**	143	−11	133	21*
Active vs. Inactive Controller	Age	−15**	169			
	Average Lecture Grade	26***	168	−14	195	24**
	Composite Instructor Rating	24***	169	−24	188	21**
	1957 Average Supervisor Rating	24***	162	−11	170	23**
With FAA vs. Not with FAA	Age	−03*	169			
	Composite Instructor Rating	16**	169	−24	188	16*
	1957 Average Supervisor Rating	20**	162	−11	170	20*
No Disciplinary Action vs. Disciplinary Action	Age	−02*	133			
	1957 Average Supervisor Rating	28***	142	−07	135	28**
	Regression Equation 2	17**	138	−06	133	17*

¹ Decimal points have been omitted.
* Significant at less than the .05 level.
**Significant at less than the .01 level.
ª Point-biserial correlations.
Taken from Trites, 1961

53.

Table 11

First Order and Partial Correlations Between the 1961 Criterion Measures and Psychological Test and Biographical Measures Where the First Order Correlations with the 1961 Criteria Are Statistically Significant

		Variable	Criterion		Age		Criterion r with Age
Criterion	No.	Description	r[1]	N	r	N	Partialed Out
1961 Average Supervisor Rating	23	Age	−23**	149			
	9	Symbolic Reasoning & Perceptual Speed (Test 5, Cal. Capacity Quest)	21**	149	−17*	135	18*
	10	Abstract Reasoning, DAT	18**	149	−26**	135	13
	11	Space Relations, DAT	18*	148	−06	135	17*
	13	Spatial Orientation, AFOQT	23**	143	−10	135	21*
	26	Previous Flying Exper.	−19**	139	67***	135	−05
	29	Any Air Traffic Exper.	18**	139	−57***	135	06
Active vs. Inactive Controller	23	Age	−15**	169			
	21	Family Relations, CTB Cal. Test of Personality	21***	150	−03	175	21**
With FAA vs. Not with FAA	23	Age	−03*	169			
	24	Education	−16**	168	44**	196	−16*
Mean Sick Leave	23	Age	−18**	142			

[1]Decimal points have been omitted.
*Significant at less than the .05 level.
**Significant at less than the .01 level.
*Point-biserial correlations.

Taken from Trites, 1961

54.

Expansion of the ATC Research Battery by Cobb

During the period that Trites was collecting the criterion data for
the five-year follow-up of the sample tested by Brokaw, Cobb began collecting
data on a test battery composed of measures that would be available to the
Federal Aviation Agency for operational use (Cobb, 1962).

Brokaw (1957, 1959) had found that certain tests administered in 1956
were predictive of both school and on-the-job criteria; Trites (1961) demon-
strated the enduring validity of those measures for significant periods of
performance on the job. The need for the research initiated in 1960 becomes
apparent when the following issues, cited by Cobb (1962) are considered:

"(1) Differences in method and content between the 1956 and the 1960-
61 training course.

"(2) The possibility of differences between aptitude levels for the
1956 and the 1960-61 training groups.

"(3) The nonavailability of United States Air Force aptitude tests for
extended use with a civilian population.

"(4) The inclusion of psychological tests measuring factor areas not
previously covered by the 1956 experimental battery.

"(5) The desirability of additional evidence to substantiate the
general findings reported by Brokaw and Trites."

Cobb's project was begun in August 1960, with the administration of an
extensive battery of tests to incoming ATC students. After initial testing
the battery was revised in September 1960, and this stabilized battery was
administered to all entering students through April 21, 1961 (Cobb, 1962).
A complete listing of variables included in Cobb's analysis appears in Table
12.

Cobb used four criterion measures in his 1962 analyses: (1) The
Combined Academic-Laboratory Grade Average. This was the mean of two
composite means: (a) the arithmetic mean of all examination grades
achieved by the student at various training levels for seven academic
subjects, and (b) pass vs fail, based upon final performance grades for
laboratory-simulated air traffic control work; (2) Pass-Fail in the school.
All students who successfully completed the training course were classified
as "pass"; those who were eliminated for deficiency during training were
classified as "fail." Students who left the course for other reasons were
dropped from the study; (3) The Scaled Objective Personality Rating. Two
psychologists evaluated statements made by instructors in a personality
profile included in every passing student's final evaluation form. The
statements concerned (a) performance under stress, (b) attitudes toward
instruction, (c) ability to work with others, and (d) job interest. The
psychologists rated every statement as positive or negative and the score
was the algebraic sum of the ratings. The value assigned to each student
was the mean of the two psychologists' ratings, which were found to corre-
late highly (.92); and (4) The Scaled Subjective Personality Rating. The

55.

Table 12

Listing of Variables Included in the Analyses

(Descriptions are provided for only those psychological test variables
which regression analyses indicated as being significant in the predic-
tion of training course criteria.)

Criterion Variables:

    A. The Combined Academic-Laboratory Grade Average.
    B. The Pass-Fail Criterion.
    C. The Scaled Objective "Personality" Rating.
    D. The Scaled Subjective "Personality Rating.

Background Variables:

    1. Age When Tested.
    2. Sum of Coded Relevant Experience.
    3. Coded Educational Background.

Psychological Test Variables:

    Variables 4-8 represent subtests of the Bennett-Seashore-Wesman
    Differential Aptitude Test (DAT) Battery, Form A.

    4. DAT-Space Relations. A 45-item test of ability to visualize objects
    and forms in two or three dimensions. The task, for each item, is
    to indicate how many of five depicted solid figures can be made from
    an unfolded pattern.

    5. DAT-Numerical Ability. A 40-item-test presenting a series of rela-
    tively simple numerical problems. Provides a measure of "number"
    ability.

    6. DAT-Abstract Reasoning. A 50-item test wherein the task is to indi-
    cate, for each item, which of a series of choices (figures) properly
    carries out a principle of logical development exhibited by a sequence
    of figures. The test provides a nonverbal measure of reasoning.

    7. DAT-Language Usage, Part II.

    8. DAT-Mechanical Reasoning.

    Variables 9 and 10 represent Part I and Part II of the Air Traffic
    Problems Test which was developed under contractural arrangement in
    1952 by the American Institute for Research for the Civil Aeronautics
    Administration.

    9. Air Traffic Problems, Part I. A 30-item test presenting highly sim-
    plified versions of Air Traffic Control situations. Good performance
    is not necessarily dependent on past ATC experience. Flight data

displays are presented for several inbound aircraft, all flying
the same speed and course, but at different altitudes and with
different ETA's. Given a basic 5-minute time separation rule,
the examinee must decide, for each item, whether or not sufficient
time separation exists between certain aircraft to permit changes
to certain specified altitudes.

10. Air Traffic Problems, Part II.

Variables 11-28 represent 18 scales of the 480-item California
Psychological Inventory (CPI) booklet. The scales provide a com-
prehensive survey of the individual from a social interaction
viewpoint, and are referred to below in terms of the factors
measured.

11. CPI-Ac (Achievement via Conformance).

12. CPI-Ai (Achievement via Independence).

13. CPI-Cm (Communality).

14. CPI-Cs (Capacity for Status). An index of an individual's
capacity for status (not his actual or achieved status). The
scale attempts to measure the personal qualities and attributes
which underlie and lead to status.

15. CPI-Do (Dominance).

16. CPI-Fe (Femininity).

17. CPI-Fx (Flexibility).

18. CPI-Gi (Good Impression).

19. CPI-Ie (Intellectual Efficiency). The degree of personal and
intellectual efficiency which a person has attained.

20. CPI-Py (Psychological Mindedness). The degree to which the indi-
vidual is interested in, and responsible to, the inner needs,
motives, and experiences of others.

21. CPI-Re (Responsibility).

22. CPI-Sa (Self Acceptance).

23. CPI-Sc (Self Control).

24. CPI-So (Socialization).

25. CPI-SP (Social Presence).

26. CPI-Sy (Sociability). Outgoing, sociable, participative temperament.

27. CPI-To (Tolerance).

28. CPI-Wb (Sense of Well Being). A scale identifying persons who minimize their worries and complaints, and who are relatively free from self-doubt and disillusionment.

   Variables 29-40 pertain to twelve subtests of the California Test of Mental Maturity (CTMM, Advanced Form A, 1957 edition).

29. CTMM-Immediate Recall.

30. CTMM-Delayed Recall.

31. CTMM-Sensing Right and Left.

32. CTMM-Manipulation of Areas.

33. CTMM-Opposites.

34. CTMM-Similarities.

35. CTMM-Analogies. A 15-item test, wherein seven drawings of different objects are presented for each item. The first object has a definite relationship to the second which the student must recognize in order to identify, by analogy, the drawing among the last four which is similarly related to the third drawing.

36. CTMM-Inference. A 15-item test, wherein printed statements for each item present two premises. The student must select the logical conclusion, based on those premises, from the four possible alternatives offered.

37. CTMM-Number Series.

38. CTMM-Numerical Quantity, Coins.

39. CTMM-Numerical Quantity, Arithmetic.

40. CTMM-Verbal Concepts.

   Variables 41-47 are representative of seven tests of the Moran Repetitive Measurements (RPM) battery. The battery is composed of highly-speeded perceptual, coordination, and memory tests. All RPM scores used in the present study were measures of performance representing initial administration.

41. RPM-A, Aiming. This test measures the ability to carry out quickly and precisely a series of movements requiring eye-hand coordination. Specifically, the student's task is to place a stylus point through the center of randomly positioned printed circles of .08-inch diameter.

42. RPM-FC, Flexibility of Closure.

43. RPM-NF, Numerical Facility.

58.

44. RPM-PS, Perceptual Speed.

45. RPM-SC, Speed of Closure.

46. RPM-V, Visualization.

47. RPM-SM, Social Memory. This test measures the ability of a student to remember faces or photographs. After studying a group of 16 photographs (faces) for one minute, the student must turn a second sheet and indicate recognition of the 16 faces from among a group of 32 pictures. Only 16 of these faces are the same. The three parts that compose the test are similar but different photographs are involved in each part.

Taken from Cobb, 1962

59.

same statements categorized in the Objective Personality Rating were evaluated on a normalized nine-point scale. The ratings of the two psychologists, correlating .90, were averaged to determine the final value for each student.

Background variables for the analysis included age, to the nearest birthday, and a sum of related experience, coded on a nine-point scale for each of ten types of experience, including three types pertaining to communications, six to air traffic control, and one to ground control intercept. Educational background was coded on a nine-point scale, such that a high school non-graduate was coded 1, and an individual with six years of college was coded 9.

Cobb (1962) described his test battery as heterogeneous, consisting mainly of commercially developed aptitude, attitude, and perceptual ability tests. Some were highly speeded and others were power tests. The selection of tests for this battery was based on a number of considerations. Some of the tests were those which Brokaw had reported as being highly predictive of training-course and on-the-job performance; several represented substitutions for USAF tests; and others were included on the assumption that they provided more comprehensive and reliable measures of certain areas or because they provided measures relevant to new or different areas.

Three samples of subjects were included in Cobb's analysis (1962):

"Sample 1, or Experimental Sample. One hundred twenty-four cases (95 pass and 29 fail subjects) from five 1960 EnRoute classes were designated as an experimental sample and scheduled for analyses aimed at the development of criterion prediction equations.

"Sample 2, or Validation Sample. A second group, composed of 172 cases (136 pass and 36 fail subjects) from eight 1961 EnRoute classes, was established as a validation sample on which to test the prediction equations derived from the analysis of the experimental sample.

"Sample 3, or Terminal Tryout Sample. One hundred forty-eight cases (137 pass and 11 fail subjects) representing thirteen 1960-61 Terminal classes, constituted a sample on which the prediction equations developed for the EnRoute classes could be tested for appropriateness in forecasting performance criteria for the Terminal course."

The intercorrelations of the four criteria in these three analytic samples are presented in Table 13. It is apparent that the academic-laboratory criterion was not highly related to the personality criteria, but that the personality criteria were highly related to each other.

Seven regression analyses were accomplished on data of the experimental sample. The first four were based on data of the combined pass plus fail cases (N=124) and resulted in the development of two prediction equations for the academic-laboratory criterion, and two for the prediction

Table 13

Intercorrelations of Criteria

Sample and Variable		Acad. Lab.	Pass-Fail	Obj. Pers.	Subj. Pers.
Experimental Sample					
Pass Plus Fails:	Acad-Lab.	1.00	.68		
Passes Only:	Acad-Lab.	1.00		.30	.28
	Obj. Pers.	.30		1.00	.80
	Subj. Pers.	.28		.80	1.00
Validation Sample					
Pass Plus Fails:	Acad-Lab.	1.00	.71		
Passes Only:	Acad-Lab.	1.00		.36	.41
	Obj. Pers.	.36		1.00	.81
	Subj. Pers.	.41		.81	1.00
Terminal Sample					
Pass Plus Fails:	Acad-Lab.	1.00	.62		
Passes Only:	Acad-Lab.	1.00		.36	.46
	Obj. Pers.	.36		1.00	.81
	Subj. Pers.	.46		.81	1.00

Taken from Cobb, 1962

61.

of the pass-fail criterion. For each criterion, one of the two equations was based on data for the psychological tests, while the other considered the tests, as well as age, experience, and education. The other three analyses were based on the 95 pass cases, and resulted in development of prediction equations for the two personality ratings and the academic-laboratory criterion. Additional regressions, designated eight and nine, consisting of a set of five tests designed to be very similar to the space relations, abstract reasoning, and air traffic problems found valid by Brokaw and by Trites, were computed for the experimental sample.

The results of Cobb's analysis of the comparative contributions of psychological and experience variables are presented in Table 14. It should be noted that regressions 1 and 3 are based upon the 95 pass cases of the experimental samples, while regressions 2 and 4 are based upon all 124 cases in the sample. The analytic technique was the iterative method described by Greenberger and Ward (1956), which sequentially selects that variable from the available set which makes the largest contribution to the derived multiple correlation. Under this system only one of the experience variables was chosen for each criterion, suggesting that available valid variance in those measures had been captured mainly by the psychological tests.

In light of the small sample sizes used in the analysis of a large number of predictor variables the relatively small shrinkage found when the composites were applied to the validation sample suggested that efficient prediction would be obtained in an unselected sample.

Comparison of the prediction of the personality ratings with the prediction of the academic-laboratory criterion appears in Table 15. Although the personality ratings were not as predictable as the performance criterion, validities at useful levels were found for them. Although substantial shrinkage occurred in the cross validation, useful prediction could be made with the personality ratings.

Of particular interest were the validities replicating the Brokaw and the Trites studies, appearing in Table 16. The observed values closely approximated the optimal prediction possible within the test battery, and cross-validation shrinkage was somewhat less.

Experience as a Predictor of ATC Performance

Following World War II, and in the period following the Korean conflict, the Federal Aviation Agency based its selection of air traffic control personnel primarily upon prior experience--either in the air traffic control field, or as a pilot. Significant numbers of persons who had relevant military experience were seeking employment. Brokaw (1959), Trites (1961), Trites and Cobb (1963), and Cobb (1962) had all demonstrated that prior experience was relevant to performance in training, although Trites (1961) could find no continuing impact in the prediction of job proficiency. In 1963 Trites and Cobb reported a more comprehensive study of the validity of experience as a predictor of training and job performance.

The experience variables evaluated and the aptitude test battery included in the study are described in Table 17, extracted from the Trites and

62.

Table 14

Development and Application of Regression Equations for Prediction of Academic Laboratory Grade Average and Pass-Fail Status of ATC En Route Students

Variables	Multiple Correlations and Beta Weights Derived Via Regression Analyses of Data for Exp. Sample							
	When Only Psych. Tests Are Considered				When Psych. Tests Are Supplemented by Back-Ground Variables			
	Regr. No. 1		Regr. No. 2		Regr. No. 3		Regr. No. 4	
	Beta Wt.	Valid-ity*	Beta Wt.	Valid-ity*	Beta Wt.	Valid-ity*	Beta Wt.	Valid-ity*
Criterion:								
A. Acad.-Lab. Grade Avg.	R=.66				R=.69			
B. Pass-Fail Status			R=.52				R=.50	
Tests:								
5. DAT-Numerical Ability					.17	.41		
6. DAT-Abstract Reasoning	.33	.54	.33	.40	.31	.54	.32	.40
9. Air Traffic, Part I	.22	.40			.16	.40		
14. CPI-Cs, Capac. For Status			-.21	-.03				
19. CPI-Ie, Intell. Eff.	.25	.29			.21	.29		
26. CPI-Sy, Sociability	-.24	-.04			-.21	-.04		
28. CPI-Wb, Sense Well Being			.19	.20				
35. CTMM-Analogies			.21	.35			.21	.35
36. CTMM-Inference	.19	.42						
41. RPM-A, Aiming	-.14	-.10	-.17	-.15	-.15	-.10		
Background:								
1. Age When Tested					-.26	-.38		
2. Coded Sum of Experience							.21	.21

	Validation Sample*: Correlation of Predicted Criterion Values with Actual			
Criterion:				
A. Acad.-Lab. Grade Avg.	.49		.56	
B. Pass-Fail Status		.35		.38

ªEXP. Sample: N=124, 95 pass plus 29 fail subjects of five 1960 En Route Classes.

ᵇVal. Sample: N=172, 136 pass plus 36 fail subjects of eight 1961 En Route Classes.

*Validity coefficients greater than .16 are significant at the .05 level and those greater than .21 are significant at the .01 level.

Taken from Cobb, 1962

Table 15

Development and Application of Regression Equations for Prediction of Criteria for Only the Pass Subjects of the ATC En Route Course

Variables	Multiple Correlations and Beta Weights Derived Via Regression Analyses of Data for Exp. Sample[a]					
	Regr. No. 5		Regr. No. 6		Regr. No. 7	
Criterion:						
A. Academic-Laboratory Grade Avg.	R = .62					
C. Objective Personality Rating			R = .47			
D. Subjective Personality Rating					R = .43	
	Beta Wt.	Valid- ity[*]	Beta Wt.	Valid- ity[*]	Beta Wt.	Valid- ity[*]
Test Variables:						
4. DAT-Space Relations	.20	.42				
6. DAT-Abstract Reasoning			.16	.21		
19. CPI-Ie, Intellectual Eff.	.33	.39	.17	.15		
20. CPI-Py, Psych. Mindedness					.14	.09
35. CTMM-Analogies			−.16	−.10		
36. CTMM-Inference	.23	.33			−.16	−.07
47. RPM-SM, Social Memory			.24	.19	.29	.22
Background Variables:						
1. Age When Tested	−.25	−.36	.15	−.00		
3. Educational Background			−.36	−.29	−.36	−.29

Criterion Variables:	Validation Sample[b]: Correlation of Predicted Criterion Values with Actual		
A. Academic-Laboratory Grade Avg.	.37		
B. Objective Personality Rating		.28	
C. Subjective Personality Rating			.20

[a]Exp. Sample: 95 pass students of five 1960 En Route classes.

[b]Val. Sample: 136 pass students of eight 1961 En Route classes.

[*]Validity coefficients exceeding .19 are significant at the .05 level and those exceeding .26 are significant at the .01 level.

Taken from Cobb, 1962

64.

Table 16

Development and Application of Regression Equations Based on Consideration of Data For Only Five Selected Tests

Variables	Multiple Correlations and Beta Weights Derived Via Regression Analyses of Data for Exp. Sample[a]			
	Regr. No. 8		Regr. No. 9	
Criterion:				
A. Academic-Laboratory Grade Avg.	R=.60			
B. Pass-Fail Status			R=.47	
	Beta Wt.	Valid-ity[c]	Beta Wt.	Valid-ity[c]
Test:				
4. DAT-Space Relations	.03	.35	−.08	.22
5. DAT-Numerical Ability	.13	.43	.12	.32
6. DAT-Abstract Reasoning	.34	.54	.27	.40
9. Air Traffic, Part I	.19	.40	.05	.23
35. CTMM-Analogies	.11	.36	.22	.35
	Validation Sample[b]: Correlation of Predicted Criterion Values With Actual			
Criterion:				
A. Academic-Laboratory Grade Avg.	.54			
B. Pass-Fail Status			.40	

[a]Exp. Sample: N=124; 95 pass plus 29 fail subjects of five 1960 En Route Classes. .

[b]Val. Sample: N=172; 136 pass plus 36 fail subjects of eight 1961 En Route Classes.

[c]Validity coefficients greater than .16 are significant at the .05 level and those greater than .21 are significant at the .01 level.

Taken from Cobb (1962)

65.

Table 17

Comparison of Experience and Aptitude Measures

Description of Variables

Variable Name & Abbreviation*	Description and Coding
Experience Variables	**Coding for Variables 1 & 2**

1. Pilot Experience (Pil)

2. Radio Operator: Air to Ground Communications (Radio: Air/Grnd)

Amount of Experience	Codes
No Experience Reported	1
Less than 1 Year	2
12 Months Through 23 Months	3
2 Years Through 4 Years	4
5 Years Through 6 Years	5
7 Years Through 8 Years	6
9 Years Through 10 Years	7
11 Years Through 15 Years	8
16 Years or More	9

3. Ground Control Intercept (GCI) — Long Range, High Altitude Interception Technique Used by Military Aircraft Control and Warning Units

4. Station (Stat) — A Unit Primarily Engaged in Ground to Air Communictions and Pilot Briefings

5. Ground to Air Communications (Ground/Air)

6. Point to Point Communications (P to P) — Communications from One Fixed Ground Point to Another Fixed Ground Point

7. VFR Tower (VFR Tow) — A Tower Controlling Air Traffic Under Visual Flight Rules (VFR)

8. Approach Control: Tower (App Con Tow) — A Tower Capable of Controlling Air Traffic Under Instrument Flight Rules (IFR) but without Access to Radar

9. Radar Approach Control: Tower (Rad App Con Tow) — A Tower with Access to Radar as an Aid in Controlling Air Traffic

10. Center (Cent) — An Air Route Traffic Control Center

11. Ground Controller Approach (GCA) — A Ground Radar System Used to Assist Aircraft During Landing

12. Radar Approach Control Center (RAPCON) — USAF Radar System Used for Approach Control at Air Force Airfields; Similar System in Use by the Navy is Called a Radar Air Traffic Control Center (RATCC)

Coding for Variables 3 Through 12

Amount of Experience	Codes
No Experience Reported	1
Through 3 Months	2
4 Through 6 Months	3
7 Months Through 1 Year	4
13 Months Through 2 Years	5
25 Months Through 3 Years	6
37 Months Through 5 Years	7
6 Years Through 10 Years	8
11 Years or More	9

13. Sum of Communications Experience (Σ Comm) — Sum of Individual Experience Variables Nos. 4, 5, & 6

14. Sum of Air Traffic Experience (Σ AT) — Sum of Individual Experience Variables Nos. 7 Through 12

15. Sum of Relevant Experience (Σ Rel) — Sum of Experience Variables Nos. 3, 13, and 14

Demographic Variables

16. Age — Chronological Age to Nearest Birthday on Date of Entry into ATCS Training

17. Education (Educ)

Coding for Education

Amount of Education	Codes
None Reported	Blank
Less than High School Graduate	1
High School Graduate	2
Less than 1 Year of College	3
1 Year of College	4
2 Years of College	5
3 Years of College	6
4 Years of College	7
5 Years of College	8
6 or More Years of College	9

66.

Table 17 Continued

Aptitude Variables

#18. DAT Space Relations (Sp) (Weight = 6)

Identify Solid Figures than Can Be Made from an Unfolded Pattern (40 Items; Scored; Rights Minus Wrongs)

#19. DAT Numerical Ability (Num) (Weight = 13)

Test of Arithmetic or Computational Skill (40 Items; Scored; Rights Minus ¼ Wrongs)

#20. DAT Abstract Reasoning (Abs) (Weight = 14)

Indicate which of a Series of Choices (Figures) Properly Carries Out a Principle of Logical Development Exhibited by a Sequence of Figures (50 Items; Scored; Rights Minus ¼ Wrongs)

##21. CTMM Analogies (Analog) (Weight = 49)

Seven Drawings of Different Objects Are Presented for Each Item. The First Object Has a Definite Relationship to the Second Which Must be Recognized in Order to Identify, by Analogy, the Drawing Among the Last Four which is Similarly Related to the Third Drawing (15 Items; Scored: Rights)

**22. Air Traffic Problems (ATP) (Weight = 18)

Determine Whether Aircraft May Be Permitted to Change Altitude without violating a Specified Time-Separation Rule (30 Items; Scored: Rights Minus Wrongs)

23. Composite Aptitude Test Score (Comp)

Sum of Aptitude Test Variables Nos. 18 Through 22, Each Weighted as Indicated.

---

* Abbreviations are given in parenthesis following the names of the variables.

Part of the Differential Aptitude Test (DAT) Battery, Form A, 1947, published by the *Psychological Corporation*, New York, N.Y.

Part of the California Test of Mental Maturity (CTMM), Advanced, Form A, 1957 published by the *California Test Bureau*, Los Angeles, California

** Originally developed by the *American Institute for Research*, Pittsburg, Pa., under contract with the Civil Aeronautics Administration in 1950. Form used in the present research was an extensive revision of the original test. Revision prepared by the *Selection Section*, Psychology Branch, Civil Aeromedical Research Institute.

Taken from Trites and Cobb, 1963

67.

Cobb report (1963). The training criteria included Academic Grade Average
(Acad.), Laboratory Grade Average (Lab.), Combined Academic plus Laboratory
Grade Average (A+L), and Pass or Fail (P-F) in the training course. The
job performance criterion variable was a supervisor rating collected approxi-
mately ten months after the trainees had graduated from the FAA Academy.
The value used was an average based upon ratings by supervisors on a 15-
item checklist evaluation form. The form contained items relating to work
habits, ability, judgment, reasoning, emotional stability, and relationships
with others. Possible ratings were: Excellent, Very Good, Good, Fair, and
Unsatisfactory. The average rating was computed by assigning weights of 4
through 0 to ratings of Excellent through Unsatisfactory, respectively,
summing all of the items rated by all supervisors, and dividing the sum by
the number of items rated. It was not possible to secure four supervisory
ratings on all subjects; the average number of forms completed was 3.8.

In an earlier study, a corrected split-half reliability of .75 was
obtained for Average Supervisor Ratings computed from rating forms con-
taining 12 of the 15 items used in the present form (Trites & Cobb, 1963).

The analytic sample included 745 trainees, of whom 505 were assigned to
EnRoute control stations and 240 Terminal assignees had participated in an
experimental aptitude testing program. Statistical comparisons of the test
samples with the total samples showed equivalence in age, academic grade,
laboratory grade, and supervisor rating (Trites & Cobb, 1963).

Trites and Cobb concluded that different kinds of pre-employment job-
related experience had differential value for the prediction of training
performance. In general, experience most directly related to air traffic
control work was a positive predictor; experience relating to communications
and piloting was negative. It was also shown that for EnRoute trainees only,
a composite variable representing the sum of tower, GCA, RAPCON/RATCC, and
center experience had a statistically significant, but low relationship with
ratings of job performance. By contrast, aptitude tests were superior to
the experience variables for the prediction of all training course performance
measures of both types of trainees, with the exception of the Laboratory per-
formance of Terminal trainees. For the Terminal trainees, the composite ex-
perience variable was superior to the tests. For job performance, it was
found that only ratings of EnRoute trainees could be predicted by the tests
and that although the tests were superior to experience as predictors, the
relationships were small. As determined in other studies it was again shown
that age at entry into training was negatively related to training and job
performance, and that the negative relationships were greatest with grades
in simulated air traffic control work in the training school laboratories.
The findings of the study led to the recommendation that selection of indi-
viduals for Air Traffic Control Specialist training be based upon an aptitude
test battery and pre-employment, job-related experience in tower, GCA, RAPCON/
RATCC, and center work.

The statistical data upon which these conclusions are based appear in
Appendix B.

Conclusions

There is a common trend across these studies for measures of spatial perception, verbal and non-verbal reasoning, and the mental manipulation of verbal or numeric concepts to be predictive of success in air traffic control training and in air traffic control assignments. Measures of prior experience in air-traffic related occupations tend to be predictive of success in training but to lose their predictive power when related to later job performance. Measures of personality, or of temperament, have shown little predictive efficiency for criteria of school or job success.

Significant difficulties were encountered in the collection of criterion data. Early studies were nearly useless because of the small groups of persons with homogeneous experience available for criterion purposes. Although later studies generated larger criterion samples, it is probable that subsets of subjects within those samples worked in dramatically different environments. The consistent findings of significant correlation in the presence of attenuation of common variance from this cause suggests that the predictor sets were reasonably powerful in their action.

The strong, stable relationship discovered by Trites between supervisory ratings of job performance and instructor ratings collected five years earlier suggests that the environment in air traffic control training and work situations is sufficiently structured to permit objective ratings of pertinent behaviors.

A crucial element in the accomplishment of air traffic control procedures lies in the efficiency of communication between controllers and aircrew members. This aspect of the controller function has not been addressed in the research reported here--*restriction of test content to materials presentable in a paper-and-pencil group-test format* effectively eliminates opportunity to measure the trainee's ability to understand what he or she may hear or the efficiency with which he or she transmits control information to the aircraft. Short-term and long-term memory in the context of the work situation also are effectively barred.

The reported research has shown that air traffic control performance can be predicted, in part, by aptitude tests in recognized areas. It has also shown that assessment of performance through classic subjective ratings can be reasonably objective and stable over time. The absence of attention to behaviors that seem to be crucial elements of the controller task indicates significant opportunity for the application of more sophisticated testing techniques. Advances in computer technology with accompanying reductions in cost will make the testing feasible in situations that simulate the work environment more accurately.

REFERENCE NOTE

1. John C. Flanagan supervised the AIR contract effort, with the advice of Elmer D. West, and Miss Marion Shaycoft. Paul M. Fitts, Jr., served as research adviser and rendered useful criticism throughout the project. L. Dewey Anderson and Bryce Hartman were consultants at early stages of the project.

Table A-1

Listing of Experimental Tests and Derived Validities of Air Force/Civil Aeronautics Administration Battery. Sample: 130 CAA Air Traffic Control Students. Extracted from Brokaw, 1957.

1. Average Lecture Grade
2. Average Instructor Rating (Pooled Independent Ratings)
3. Composite Instructor Rating (Instructors as Group Rate Students)

Test Title	Mean	SD	Validity* 1	2	3
PUBLICATIONS OF THE CALIFORNIA TEST BUREAU					
Survey of Space Relations Ability	62.54	16.64	.06	.07	.08
Occupational Interests Inventory					
Personal-Social	17.52	5.27	−.02	−0.1	.01
Natural	20.85	6.78	−.03	.01	.04
Mechanical	20.98	5.56	.13	.09	.07
Business	20.85	5.85	.08	−.03	−0.5
The Arts	16.72	5.71	−.23	−.12	−.12
The Sciences	22.94	4.83	.08	.01	.03
Types of Interest					
Verbal	9.52	3.52	−.09	−.10	−.08
Manipulative	10.58	1.69	−.07	−.03	−.07
Computational	9.73	3.51	.06	.03	.01
Level of Interest	73.56	7.03	−.01	−.06	−.02
Mental Health Analysis					
Liabilities					
Behavioral Immaturity	14.77	2.99	.13	.07	.03
Emotional Instability	15.58	3.30	.07	.05	.04
Feelings of Inadequacy	16.79	2.33	.14	.02	−.01
Physical Defects	19.57	.92	.14	.09	.09
Nervous Manifestations	18.02	2.00	.25	.08	.10
Total Liabilities	84.75	8.76	.18	.06	.04
Assets					
Close Personal Relations	18.32	2.04	−.14	−.08	−.07
Interpersonal Skills	17.50	2.17	−.08	−.08	−.07
Social Participation	14.47	3.77	−.05	−.03	−.05
Satisfying Work & Recreation	15.65	2.50	−.03	.00	.00
Adequate Outlook and Goals	18.58	1.58	−.05	−.06	−.08
Total Assets	84.52	9.41	−.09	−.06	−.06

*.17 significant at 5-per-cent level, .23 at 1-per-cent level.

Test Title	Mean	SD	Validity*		
			1	2	3
Publications of the California Test Bureau (Continued)					
Test of Mental Maturity					
Sensing Right and Left	17.72	2.06	.09	.06	.08
Manipulation of Areas	8.90	2.36	.08	−.02	.06
Similarities	7.01	2.05	.11	.00	.07
Inferences	13.16	1.44	.19	−.01	.04
Number Series	7.52	3.44	.21	.18	.23
Numerical Quantity	10.42	2.04	.28	.21	.24
Verbal Concepts	37.26	5.64	.09	−.01	.06
Survey of Working Speed and Accuracy					
Number Checking	120.52	22.01	.09	.16	.16
Code Translation	130.90	28.10	.27	.24	.27
Finger Dexterity	82.98	16.58	.00	−.03	−.02
Counting	39.65	10.01	.29	.24	.19
California Capacity Questionnaire					
Test 1 (Right or Left)	17.14	2.32	.07	.00	.03
Test 2 (Vocabulary, Arithmetic Reasoning and Syllogisms)	13.85	1.11	.11	.03	.10
Test 3 (Symbolic Reasoning, Perception)	10.92	2.10	.24	.02	.04
Test 4 (Vocabulary, Arithmetic.. Reasoning, Syllogisms)	12.36	2.02	.09	.04	.12
Test 5 (Symbolic Reasoning, Perception)	7.22	2.30	.33	.29	.28
Test 6 (Vocabulary, Arithmetic Reasoning, Syllogisms)	6.85	3.18	.17	.22	.24
Test 7 (Vocabulary)	3.78	3.56	.18	.18	.18
California Test of Personality					
Personal Adjustment					
Self-Reliance	12.38	1.94	−.02	−.09	−.07
Sense of Personal Worth	13.17	1.82	.03	.03	.02
Sense of Personal Freedom	13.61	1.61	−.08	−.06	−.06
Feeling of Belonging	13.95	1.71	.08	.02	.02
Withdrawing Tendencies	13.18	2.44	.04	−.05	−.07
Nervous Symptoms	13.58	1.77	.08	.07	.06
Social Adjustment					
Social Standards	12.62	1.66	.01	−.03	−.01
Social Skills	12.10	2.03	−.01	−.12	−.12
Anti-Social Tendencies	13.93	1.69	.03	−.05	−.05
Family Relations	13.92	1.70	.09	.19	.15

*.17 significant at 5-per-cent level, .23 significant at 1-per-cent level.

71.

Test Title	Mean	SD	Validity* 1	2	3
Publications of the California Test Bureau (Continued)					
Occupation Relations	13.11	1.78	.16	.06	.01
Community Relations	13.19	2.20	.07	−.01	−.01
Personnel Selection and Classification Test					
Reading	36.12	3.47	.28	.16	.16
Arithmetic	18.49	4.29	.33	.28	.32
Vocabulary	37.70	5.30	.15	.06	.14
Mechanical Dexterity					
Dotting	21.01	4.17	.13	−.04	−.02
Pursuit	23.65	5.93	.01	−.07	−.09
Blocks (Counting)	37.42	13.80	.12	.08	.10
Total Mechanical Factors	82.25	18.21	.12	.03	.04
Social Adaptability	24.46	3.30	−.02	−.02	−.09
Dependability	24.87	2.92	.06	.00	.01
Oral Directions	21.99	2.35	.06	.11	.13

DIFFERENTIAL APTITUDE TEST BATTERY PUBLISHED BY THE PSYCHOLOGICAL CORP.

Test Title	Mean	SD	1	2	3
Verbal Reasoning	33.85	7.96	.21	.12	.17
Numerical Ability	26.66	6.79	.32	.28	.31
Abstract Reasoning	35.74	6.61	.18	.15	.20
Space Relations	67.65	15.24	.21	.15	.20
Mechanical Reasoning	51.55	7.44	.10	−.02	.01
Clerical Speed and Accuracy	63.78	15.12	.00	.00	.01
Language Usage					
Spelling	67.03	20.93	.17	.05	.10
Sentences	41.82	13.19	.22	.18	.15

AIRWAYS OPERATIONS APTITUDE TEST BATTERY-AMERICAN INSTITUTE FOR RESEARCH

Test Title	Mean	SD	1	2	3
Air Traffic Problems I	18.88	5.23	.30	.31	.37
Air Traffic Problems II	15.33	3.96	.26	.15	.18
Locating Data	31.87	8.57	−.02	.25	.26

UNITED STATES AIR FORCE TESTS ADMINISTERED EXPERIMENTALLY

Test Title	Mean	SD	1	2	3
Airman Classification Battery AC-2A Booklet 2					
Arithmetic Reasoning	7.45	1.51	.38	.27	.26
Verbal Test	7.38	1.54	.21	.05	.13

*.17 significant at 5-per-cent level, .23 significant at 1-per-cent level.

Test Title	Mean	SD	Validity*		
			1	2	3

United States Air Force Tests Administered Experimentally (Continued)

Test Title	Mean	SD	1	2	3
Mechanical Test	6.75	1.75	.24	.00	.01
Tool Functions	6.08	1.92	.24	-.01	.01
Figure Recognition	6.16	1.98	.19	.07	.07
Airman Classification Battery AC-2A Booklet					
Technical Information	7.38	1.34	.30	-.01	.03
Patterns	5.97	1.82	.20	.14	.17
Dial and Table Reading BP622-621A	7.49	1.29	.43	.38	.38
Air Force Officer Qualifying Test Booklet 2					
Aerial Landmarks	25.84	7.64	.16	.27	.27
Spatial Orientation	4.15	2.23	.13	.22	.18
Air Force Officer Qualifying Test Booklet 3					
Instrument Comprehension	17.76	5.65	.26	.23	.23
Aerial Orientation	12.15	5.68	.16	.08	.12
Visualization of Maneuvers	10.57	5.24	.11	.14	.14
Criterion 1, Average Lecture Grade	93.56	2.63		.50	.51
Criterion 2, Average Instructor Rating	8.65	4.62			.93
Criterion 3, Composite Instructor Rating	8.92	5.11			

*.17 significant at the 5-per-cent level, .23 significant at the 1-per-cent level.

73.

Table A-2

Description of Suggested Air Traffic Control Selection Test

General

The Test Should Be Composed of Materials Appropriate for Use with Juniors and Seniors in High School. Test Directions and Explanatory Material Should Be Suitable for Use in Junior High School Because There Is Evidence that Verbal Skill, as Such, Is Not Particularly Valid as a Selection Device for this School. The Content of the Technical Portions of the Test, However, Should Be at the Senior Level in High School.

Test Outline

	No. of Items	Time Limit
Arithmetic Reasoning	30	25 Minutes
Figure Analogies	30	15 Minutes
or		
Symbolic Reasoning of the type of Test 5, California Capacity Questionaire	30	7 Minutes
Air Traffic Problems	30	5 Minutes
Code Translation	70 Word Passage	5 Minutes

Passing of Materials and Administrative Directions Will Require an Additional Ten to Fifteen Minutes, Making the Total Test Time Somewhere Between an Hour and an Hour and Ten Minutes.

Scores Derived from the Four Item Types Employed Should Be Equated by Conversion to Normalized Standard Scores and Combined with the Following Suggested Weights:

Arithmetic Reasoning 3
Symbolic Reasoning 2
Air Traffic Problems 3
Code Substitution 1

74.

Table B-1

Comparison of Certain Characteristics of the Total and the Aptitude-Tested Parts of the Enroute and Terminal Samples

	Enroute						Terminal					
	Tested			Total			Tested			Total		
Variables	X̄	S.D.	N.	X̄	S.D.	N.	X̄	S.D.	N.	X̄	S.D.	N.
Age	28.5	6.3	470	28.5	6.4	501	27.8	5.9	212	27.9	5.9	241
Academic Gd.	83.6	7.4	470	83.6	7.3	502	87.0	5.3	211	87.1	5.3	241
Lab Gd.	78.1	13.0	467	78.0	13.0	499	83.3	5.2	207	83.3	5.4	237
A + L Gd.	80.8	9.3	468	80.8	9.2	499	85.0	5.6	211	85.0	5.6	241
Supervis. Rating	2.45	.54	301	2.45	.53	318	2.42	.67	180	2.43	.66	204

Table B-2

Means and Standard Deviations of the Experience Variables, Test Variables, and Education Based on the Maximum Possible Number of Trainees in the Enroute and Terminal Samples Separately

	Mean		Std. Dev.		N	
Variables	E	T	E	T	E	T
1. Pilot	2.38	1.92	2.64	2.16	503	242
2. Radio: Air/Grnd	1.38	1.28	1.23	1.13	503	242
3. GCI	1.39	1.23	1.42	1.07	503	242
4. Station	1.39	1.10	1.38	.68	503	242
5. Grnd/Air	2.12	1.72	2.22	1.88	503	242
6. P to P	1.39	1.23	1.34	1.08	503	242
7. VFR Tower	2.68	4.19	2.30	2.51	503	242
8. App Con Tow	1.80	2.53	1.68	2.16	503	242
9. Rad App Con Tow	1.14	1.14	.77	.72	503	242
10. Center	1.44	1.20	1.28	.78	503	242
11. GCA	2.21	2.26	2.14	2.25	503	242
12. RAPCON	1.75	1.46	1.68	1.35	503	242
13. Σ Comm	4.89	4.07	3.51	2.57	503	242
14. Σ AT	11.01	12.81	4.83	4.21	503	242
15. Σ Rel	17.28	18.10	5.06	4.46	501	242
17. Educ	2.90	2.79	1.59	1.49	489	241
18. Space	60.79	58.98	18.93	19.62	472	212
19. Numerical	22.17	21.68	8.13	7.52	474	212
20. Abstract	33.92	34.48	7.14	7.16	474	212
21. Analogies	6.80	6.71	1.99	2.03	475	212
22. ATP	13.46	13.34	5.84	6.06	475	212
23. Composite	1703.	1688.	363.	375.	471	212

Table B-3

Enroute Sample: First-Order Correlations and Partial Correlations (Age Held Constant Statistically) Between Experience Variables and Criterion Measures

Variables	A + L r	A + L Partial r	P − F¹ r	P − F¹ Partial r	Acad r	Acad Partial r	Lab r	Lab Partial r	Super r	Super Partial r	Age r
1. Pilot	−12**	07**	−17**	02	01	14**	−18**	02	−17**	−07	64**
2. Radio: Air/Grnd	−23**	−19**	−15**	−11**	−20**	−18**	−20**	−16**	03	07	18**
3. GCI	−14**	−13**	−12**	−11**	−16**	−16**	−10*	−09*	−01	00	04
4. Station	−11*	−11*	−11*	−11*	−17**	−17**	−06	−06	−02	−01	02
5. Grnd/Air	−15**	−15**	−13**	−13**	−23**	−22**	−09*	−09*	−05	−05	03
6. P to P	−12**	−12**	−15**	−15**	−21**	−21**	−05	−04	−15**	−15**	03
7. VFR Tower	14**	09*	16**	11*	21**	18**	09*	03	11*	08	−21**
8. App Con Tow	17**	16**	16**	14**	19**	18**	13**	11*	03	01	−09
9. Rad App Con Tow	06	06	06	07	08	08	04	04	02	02	01
10. Center	14**	13**	08	07	08	07	15**	14**	09	08	−05
11. GCA	13**	10*	18**	16**	14**	13**	11*	09	09	08	−10*
12. RAPCON	16**	12**	15**	11*	14**	11*	15**	10*	05	02	−17**
13. Σ Comm	−18**	−18*	−18**	−18*	−29**	−29**	−10*	−10*	−10	−09	03
14. Σ AT	29**	24**	30**	25**	31**	29**	23**	18**	16**	12**	−24**
15. Σ Rel †	11	06	13**	08	05	02	12**	07	09	06	−19**
16. Age †	−27**		−28**		−15**		−30**		−18**		
Total N	499		495		502		499		318		501

¹Point-biserial correlations; all other correlations are product-moment. Decimal points omitted from all correlations.
* Significant at less than the .05 level.
**Significant at less than the .01 level.
†For Σ Rel and Age the Ns used in the correlations with Acad and Lab were 500 and 498, respectively. For the other correlations in these columns, the Ns were as indicated in the Total N row.

Table B-4

Terminal Sample: First-Order Correlations and Partial Correlations (Age Held Constant Statistically) Between Experience Variables and Criterion Measures

Variables	A + L r	A + L Partial r	P − F¹ r	P − F¹ Partial r	Acad r	Acad Partial r	Lab r	Lab Partial r	Super r	Super Partial r	Age r
1. Pilot	−12	−02	−13*	−01	04	10	−32**	−18**	−17**	−14*	51**
2. Radio: Air/Grnd	−07	−04	03	07	−07	−06	−09	−05	−17*	−16*	15*
3. GCI	−06	−01	−04	01	−03	−01	−11	−04	01	03	21**
4. Station	−10	−10	02	03	−15*	−15*	−07	−07	01	02	03
5. Grnd/Air	−07	−08	03	02	−13*	−14*	−06	−08	−01	−02	−04
6. P to P	−08	−10	−02	−04	−09	−10	−10	−12	06	05	−04
7. VFR Tower	08	03	13*	09	−01	−03	26**	21**	02	00	−20**
8. App Con Tow	20**	20**	06	07	10	10	25**	27**	02	02	00
9. Rad App Con Tow	06	09	06	09	04	05	06	11	03	04	12
10. Center	05	04	08	05	07	06	01	−02	01	00	−09
11. GCA	04	06	01	03	01	02	03	06	−01	−01	09
12. RAPCON	11	11	03	03	09	08	11	11	06	05	−03
13. Σ Comm	−11	−12	02	01	−17**	−17**	−10	−11	02	02	−03
14. Σ AT	22**	22**	15*	14*	10	10	35**	35**	05	04	−07
15. Σ Rel	13*	13*	14*	14*	−01	−01	25**	26**	06	06	−03
16. Age †	−21**		−25**		−09		−34**		−09		
Total N	241		240		241		237		204		241

¹Point-biserial correlations; all other correlations are product-moment. Decimal points omitted from all correlations.
* Significant at less than the .05 level.
**Significant at less than the .01 level.
†Correlations between Age and the first four criteria were computed with an N which was one less than the Ns given in the Total N row.

Frequencies of Trainees Reporting Various Types of Experience and Chi Square Tests of the Experience—No Experience Dichotomy vs. Pass-Fail and Above-Below the Approximate Median of Academic plus Laboratory Grade Average

Variables	Enroute				Enroute: A + L Grade				Terminal: A + L Grade			
	Fail	Pass	Total	X²	≤83	≥84	Total	X²	≤85	≥86	Total	X²
1. Pilot	45	80 (92)#	125	8.57*	80	45 (58)	125	7.11	25	21 (25)	46	NS*
2. Radio: Air/Grnd	23	27 (37)	50	11.47	40	11 (24)	51	13.97	13	4 (9)	17	7.00
3. GCI	14	22 (27)	36	NS	27	11 (18)	38	4.98	9	3 (7)	12	4.39
4. Station	17	23 (30)	40	6.10	25	15 (19)	40	NS	5	1 (3)	6	NS
5. Grnd/Air	39	71 (81)	110	6.47	71	39 (51)	110	6.67	22	11 (18)	33	6.81
6. P to P	21	22 (32)	43	12.67	34	10 (20)	44	10.78	9	3 (7)	12	4.39
7. VFR Tower	33	166 (147)	199	15.51	88	111 (92)	199	11.98	71	92 (89)	163	NS
8. App Con Tower	11	95 (78)	106	17.21	41	63 (50)	107	12.97	30	61 (49)	91	9.47
9. Rad App Con Tow	2	16 (13)	18	NS	9	9 (8)	18	NS	4	7 (6)	11	NS
10. Center	9	50 (44)	59	4.06	20	39 (27)	59	10.56	8	9 (9)	17	NS
11. GCA	17	130 (109)	147	22.79	61	87 (69)	148	13.20	33	39 (39)	72	NS
12. RAPCON	10	86 (71)	96	15.12	29	68 (45)	97	27.45	10	21 (17)	31	NS
** Total N	129	366	495		268	231	499		110	131	241	

Numbers in parentheses are expected values for Pass or Above Median trainees computed from a 2 × 2 contingency table. If the expected value is less than the observed value, the sign of the relationship is positive; if greater, negative.
* With 1 degree of freedom a X² value of 3.84 is significant at the .05 level, 6.64 at the .01 level, and 10.83 at the .001 level. NS replaces X² values not reaching the .05 significance level. Yates correction was not applied.
**Total N is the maximum possible frequency which could occur as an entry in each column. From the Total N and the table entries, the 2 × 2 contingency tables can be reconstructed.

Enroute Sample: First-order Correlations and Partial Correlations (Age Held Constant Statistically) of Test Variables and Education with the Criterion Measures

Variables	A + L		P − F[1]		Acad		Lab		Super		Age
	r	Partial r	r	Partial r	r	Partial r	r	Partial r	r	Partial r	r
17. Educ	−03	05	−08	00	03	07	−06	03	−08	−01	28**
18. Space	34**	31**	30**	26**	37**	35**	28**	24**	04	00	−18**
19. Numerical	36**	39**	26**	30**	43**	45**	27**	30**	09	11	08
20. Abstract	45**	43**	36**	33**	44**	43**	39**	36**	12*	09	−17**
21. Analogies	28**	25**	18**	15**	30**	28**	23**	20**	13*	11*	−14**
22. ATP	37**	34**	28**	25**	32**	30**	35**	31**	15*	12*	−17**
23. Composite	52**	49**	39**	37**	53**	52**	44**	41**	17**	14*	−17**
Max. Range of N	487		483		488		486		310		489
Min.	468		464		470		467		301		470

[1]Point-biserial correlations; all other correlations are product-moment. Decimal points omitted from all correlations.
* Significant at less than the .05 level.
**Significant at less than the .01 level.

Terminal Sample: First Order Correlations and Partial Correlations (Age Held Constant Statistically) of Test Variables and Education with the Criterion Measures

Variables	A + L		P − F[1]		Acad		Lab		Super		Age
	r	Partial r	r	Partial r	r	Partial r	r	Partial r	r	Partial r	r
17. Educ	00	04	−03	01	05	06	−08	−03	01	00	16*
18. Space	14*	12	20**	17*	26**	25**	06	02	01	−01	−11
19. Numerical	33**	34**	19**	20**	45**	45**	14*	15*	10	10	01
20. Abstract	40**	37**	38**	35**	45**	44**	30**	24**	−03	−06	−25**
21. Analogies	12	10	09	08	22**	22**	02	−01	01	00	−08
22. ATP	37**	34**	33**	29**	38**	37**	31**	25**	07	06	−22**
23. Composite	38**	35**	33**	30**	49**	48**	23**	18**	05	03	−18**
Max. Range of N	240		239		240		236		203		240
Min.	211		210		211		207		180		212

[1]Point-biserial correlations; all other correlations are product-moment. Decimal points omitted from all correlations.
* Significant at less than the .05 level.
**Significant at less than the .01 level.

77.

Chapter 4

THE SELECTION OF AIR TRAFFIC CONTROL SPECIALISTS: CONTRIBUTIONS BY THE CIVIL AEROMEDICAL INSTITUTE

William E. Collins, James O. Boone, and Allan D. VanDeventer

HISTORICAL REVIEW

Eligibility for Air Traffic Control Specialist (ATCS) training with the Federal Aviation Administration (FAA) has traditionally included consideration of an applicant's preemployment experience, educational background, the outcome of an interview with management officials, and the results of a medical examination. Previous relevant experience, particularly in military air traffic control, has always been heavily weighted in the selection process. Experience as a pilot and various type of previous work in communications and air surveillance have also been consistently viewed as important assets. In general, however, ATCS selection programs prior to 1962 involved no formal assessment of mental abilities or aptitudes (Cobb, 1971).

ATCS Selection Research

Historically, research on ATCS selection is rooted in a 1950 contract by the Civil Aeronautics Administration (CAA--predecessor to the FAA) for the development of aptitude tests that could be used to select ATCS trainees. The results of that contracted study: (1) indicated that aptitude tests potentially could make an effective contribution in the selection process, and (2) provided the format for an Air Traffic Problems (ATP) test (Trites and Cobb, 1964a).

In 1956, Brokaw of the United States Air Force's Personnel Laboratory, in a joint effort with the CAA, administered a large number of aptitude tests to 197 ATCS trainees. His findings (Brokaw, 1959) indicated that a composite aptitude test score formed by adding together scores on tests of arithmetic reasoning, symbolic reasoning, code translation, and the ATP test could effectively predict instructors' ratings of training performance and supervisors' ratings of job performance approximately a year after training.

Then in 1959, a number of ATCS trainees were recruited on an experimental basis using the Federal Service Entrance Examination. For a variety of reasons, the results of that experiment were considered inconclusive (Trites and Cobb, 1964a).

A continuing research program on ATCS selection began in 1960 with the establishment of the FAA Civil Aeromedical Research Institute (now the Civil Aeromedical Institute--CAMI) in Oklahoma City, Oklahoma. The first study (Trites, 1961) was a followup of the 197 trainees tested by Brokaw in 1956. Job performance ratings were obtained by Trites from supervisors of 143 of

79.

the former trainees; the vast majority of the latter were fully qualified
ATCSs. To evaluate the effectiveness of the aptitude tests recommended
by Brokaw, the 143 men were classified as either satisfactory or marginal
in their job performance as determined from the ratings and comments of
their supervisors. The distribution of the composite scores based on the
four aptitude tests given each trainee in 1956 was divided into fourths
and the percentages of satisfactory and marginal individuals in each fourth
were computed (see Figure 1). Scores achieved on the aptitude tests given
at the beginning of training were clearly related to job performance evalua-
tions obtained 4 to 5 years after training.

Concurrently with the collection of the facility performance data on
the trainees tested in 1956, extensive experimental psychological testing
of all EnRoute and Terminal trainees who entered the FAA Academy in Oklahoma
City for basic ATCS training was begun in August 1960 (Trites and Cobb,
1964a). All of these trainees had been selected by medical and experience
requirements and none by aptitude tests. The group of commercially available
tests with which experimentation was started required 8 hours for administra-
tion. It included tests previously found most useful for prediction of ATCS
training and job performance plus tests of a number of aptitude areas not
previously examined and yielded 44 different scores. These tests were either
commercially published instruments or aptitude assessment devices developed
under contractual arrangement for the FAA; no Civil Service Commission (CSC--
now known as the Office of Personnel Management, OMP) tests were included in
the research prior to July 1961 (Cobb, 1971). By that time a series of
multiple-regression analyses, accomplished by Cobb and Trites in connection
with followup studies of several hundred men, had identified a total of 8
tests (from a group of 27) from which a variety of summary measures having
validity for prediction of ATCS trainee performance might be derived. Seven
of the eight tests were commercially published instruments; the remaining
one was the contractually developed ATP test.

While the seven published instruments of that experimental battery are
referred to as "tests," they were actually parts, taken from rather lengthy
and comprehensive aptitude-measuring devices. Three of the seven were sub-
tests of the Psychological Corporation's Differential Aptitude Test (DAT),
namely, DAT Space Relations, DAT Numerical Ability, and DAT Abstract
Reasoning. The first test measures the ability of an examinee to visualize
objects and forms in two or three dimensions, the second is a test of arith-
metical or computational skill, while the third provides a measure of non-
verbal reasoning (specifically, determining, for each item, which of a
series of choices-figures properly carries out a principle of logical
development exhibited by a sequence of figures). The remaining four
were subtests of the California Test Bureau's Test of Mental Maturity
(CTMM, Advanced Form A Edition; Sullivan, Clark, and Tiegs, 1957). In
each item of CTMM Analogies, the examinee must recognize the relationship
between a pair of drawings (objects) in order to identify, by analogy,
one of four choices as being similarly related to a third. CTMM Inference
involves the comprehension of statements that present premises underlying
the derivation of logical conclusions. The subtest designated as CTMM
Numerical Quantity-Arithmetic measures the ability to solve word-presented
arithmetic problems, while CTMM Numerical Quantity-Coins involves the mental
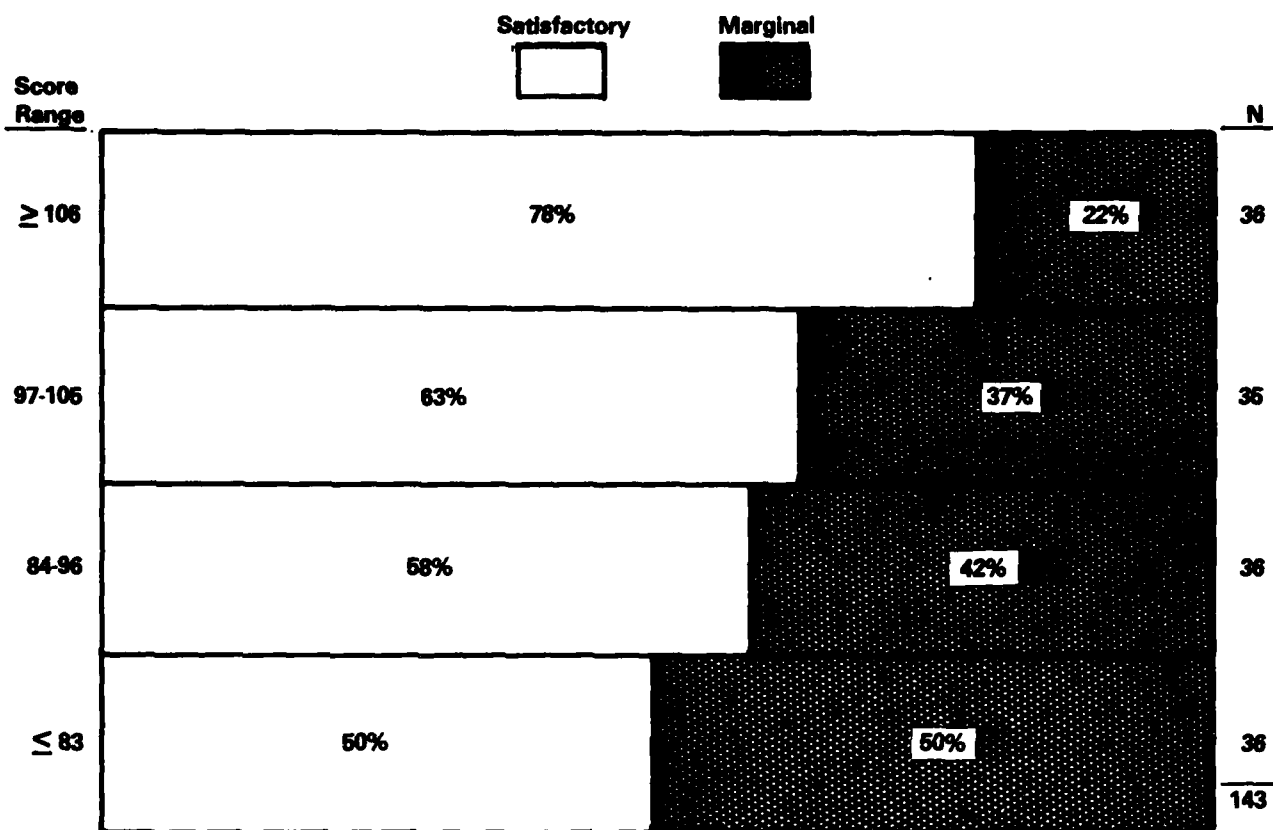manipulation of interrelated amounts of money and numbers of coins.

80.

Figure 1. Percentages of Marginal and Satisfactory ATCSs in Approximate Fourths of Composite Aptitude Test Scores Predicted from an Early Regression Equation (from Trites and Cobb, 1964).

81.

Under the time conditions available, no single group of trainees was ever administered all eight tests (the seven commercial tests plus Air Traffic Problems); one class was examined with seven, and several successive classes were administered either six or five of the eight instruments. When an average of the academic and laboratory grades was computed for each Academy trainee and employed as the criterion, the validities of the composite measures, derived from the performance scores on five, six, or seven tests, were found to range from .35 to .54 (Pearson product-moment coefficients). More importantly, an analysis revealed that about 70% to 80% of the cases classified as training-course failures were, in most instances, represented in the lower half of the distribution of scores derived with each respective group of tests. Moreover, the Academy attrition rates were averaging over 30% despite the fact that most trainees were former military controllers.

Composition of a test battery for the CSC. The potential value of the tests for screening purposes was thus recognized, and, when similar results were obtained with additional samples, the FAA and the CSC agreed that aptitude-test measures should be employed on a tentative basis in the selection of some of the nonexperienced applicants (Cobb, 1971). However, commercially published tests could not be used because such instruments were deemed more susceptible to compromise than those subjected to rigid CSC control procedures. Commission officials therefore examined their extensive file of CSC tests and selected several instruments that, in terms of factor content, appeared to approximate a number of the validated commercial tests. Since the ATP test had been developed specifically for the FAA and was still completely controlled, it was adopted officially as a CSC test. In addition, CAMI researchers were provided a number of other CSC tests to administer and evaluate.

Commencing in August 1961, all incoming classes of Academy ATCS trainees were assessed experimentally with the entire group of tests extracted from the CSC files and with the ATP test. The restricted time available for each testing session precluded examination of each class with the complete and previously validated battery of commercial tests. Nevertheless, time beyond that required for the CSC tests was available to permit administration of a portion of the commercial battery. Follow-up studies of Academy trainees examined with the revised battery during the next 10 months revealed that composite scores based on five of the CSC tests and the ATP test could be used effectively to predict training outcomes (Cobb, 1971). Composite scores of 190 and higher were attained by approximately 55% of all the examinees. Of these, about 70% successfully completed their training course and were certificated as ATCSs. In contrast, almost 75% of those with scores of 189 and lower failed to graduate and were eliminated from further FAA training. These results approximated those obtained in earlier analyses (with other groups) for the "commercial seven-test composite." Three of the tests involved in the six-test CSC composite were those that had been selected as "counterparts" of three commercial tests; they measured numerical, spatial, and nonverbal abstract-reasoning abilities. The new composite also included the ATP test (an extensive revision of the original test), an instrument known as Letter Sequence (a new test, not identified in the original research, which

82.

measured reasoning ability), and a test of following oral directions. The CSC battery was commonly referred to as "The CSC ATC Aptitude Screening Test" and its six elements as "subtests."

CAMI was requested to continue its experimental testing program and obtain additional validation data including, now, Flight Service Station (FSS) ATCS trainees for whom no previous experimental data were available. However, an early analysis, in which the aptitude test scores of the first 302 examinees were validated against the Academy training criteria, yielded findings that were highly similar to those later obtained for the complete sample and that prompted CSC and FAA officials to authorize use of the battery, beginning in July 1962, for the screening of applicants who were unable to establish training eligibility in terms of the normally pre-scribed qualification standards (i.e., qualifications with respect to aviation-related experience and/or education). Several thousand such applicants were operationally examined with the battery during the follow-ing 18 months, and although about half of them established training eli-gibility by achieving raw composite scores of 190 or higher (i.e., per-centile scores of 70 or better), very few were selected (Cobb, Young, and Rizzuti, 1976). Candidates who qualified on the basis of previous ATC work and other aviation experience generally attained higher overall CSC eligibility ratings than those screened with the battery. Moreover, training quotas continued to decline and were usually met by the selection of candidates with CSC percentile ratings of 90 or higher. In order to attain a percentile rating of 90, an applicant with insufficient ATC-related experience to qualify for any credit points was required to achieve an exceptionally high aptitude test score of 257. Aptitude-screened applicants were, therefore, seldom able to compete effectively for the available training positions. In fact, most of the relatively few aptitude-screened candidates selected for training prior to January 1964 possessed at least some ATC-related experience that, although in-sufficient for exemption from the test screening requirement, warranted credit points to supplement ratings reflecting excellent levels of per-formance on the test battery.

Academy training performance records for the last of the 893 candi-dates examined with the CSC test battery for research purposes only were not available until October 1963. By that time, Cobb and Trites had completed several sets of analyses of Academy performance (see Figure 2) and had collected post-Academy·information on training progress, experi-mental ratings of job performance, and other data for several hundred of the examinees who had successfully completed their basic training course some 12 to 18 months earlier.

A series of validation analyses completed shortly thereafter yielded findings of timely interest to officials seeking to improve ATCS selection (Cobb, Young, and Rizzuti, 1976). Perhaps the most important of the analyses was one which, in support of the preliminary findings obtained in the August 1961 to May 1962 analyses, revealed that about two-thirds of the 271 Academy attritions among the 893 experimentally examined trainees scored no higher than 189 on the CSC battery, whereas two-thirds of the 622 graduates scored 190 or higher. In another analysis based on the entire sample, statistically significant correlations were obtained
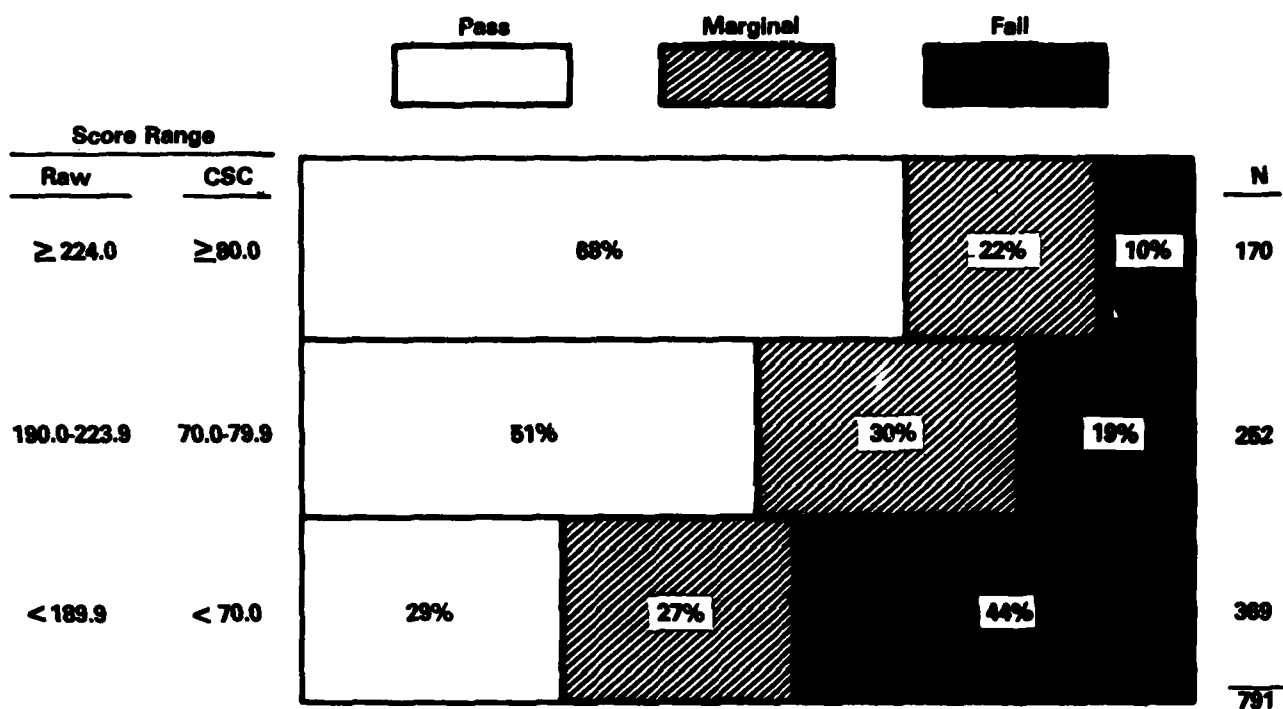
83.

Figure 2. Percentages by Three Composite Score Ranges of Passing, Marginal, and Failing Trainees Who Took the CSC ATC Aptitude Screening Test Experimentally During 1960-63: N=791 EnRoute, Terminal, and FSS Trainees Combined (from Trites and Cobb, 1964).

84.

between the aptitude test variable and most of the Academy training-performance measures (i.e., grades). However, for Academy graduates only, the aptitude-test scores (particularly those above 210) failed to predict either Academy training performance or promotions, ratings of job performance, or attrition-retention status during the first 12 to 18 months of facility training. These latter findings were not surprising in view of the restriction of range caused by the absence in this sample of data for Academy attritions, the majority of whom were low-aptitude trainees.

Adoption of the CSC Battery. These results prompted a revision in selection standards. Beginning in January 1964, the CSC battery was incorporated in the screening of all applicants regardless of pre-FAA experience. Aside from other factors, eligibility required a composite CSC score of at least 210. Retention of a screening score of 190 was considered, but 210 was eventually adopted because it was contemplated by FAA officials that a further reduction in the number of ATCS trainee positions would be necessary pending an increase in congressional appropriations.

Revised aptitude-screening procedures. The screening program instituted in January 1964 required that all applicants for ATCS training be examined with the CSC battery. However, they were screened on the basis of three different aptitude-test-performance qualification standards. In accordance with procedures prescribed for each specific training option and for the entry pay grade specified by the applicant, an individual's preemployment experience and/or level of education determined which of three tables was to be used in converting the applicant's composite raw score on the test battery to a percentile score. A percentile score of 70, was established as a mandatory eligibility requirement, but this corresponded to a raw score of 210 on one conversion table, 225 on another, and 240 on the third. For example, for candidates for Tower or Center training at the GS-6 (General Schedule pay grade) entry level, it was required that ATC-rated applicants (usually former military controllers) score at least 210 on the battery; 225 was considered minimally qualifying for instrument-rated, aircraft-pilot personnel or applicants having navigator or air-dispatcher certificates, and 240 served as the screening standard for those having low-to-moderate amounts of aviation-related experience, a 4-year college degree, or certain diverse experiential and educational backgrounds. Moreover, the procedures stipulated that each applicant's percentile score, if 70 or higher, be supplemented with credit points awarded for those types of experience that had warranted use of either of the two lower test-score screening standards, in order to derive the individual's overall eligibility rating (Cobb, Young, and Ruzzuti, 1976). Inasmuch as the majority of the applicants possessed aviation experience of some sort, the dual consideration of that experience in the qualification process enabled them to establish training candidacy in far greater numbers, and generally with higher eligibility ratings, than those with nonaviation backgrounds.

These relatively high screening standards, of 210, 225, and 240 for the respective training options remained in effect from January 1964 until August 1968, and resulted in the screenout of more than half of the applicants. However, no shortage of qualified candidates developed. Continuing

budgetary limitations kept training quotas unusually low throughout the
entire 56-month period and consequently nearly all ATCS trainees were
selected from among candidates with very high eligibility ratings--
reflecting exceptional qualifications with respect to aptitudes,
experience, and education.

   Medical examination. Eligibility for ATCS training within the FAA
has required applicants, after having met other qualifications, to pass
a rigid medical examination. The procedures relating to medical certi-
fication of ATCS personnel were amended in 1966 to include considera-
tion of personality attributes as well (Cobb and Nelson, 1974). Since
that time, Cattell's Sixteen Personality Factor (16 PF) Questionnaire
(Cattell and Éber, 1962) has been used in the screening of all applicants.
Medically qualified individuals for whom the 16 PF Questionnaire results
indicate no significant emotional or mental problems have been granted
medical certification. Others, usually a small minority of the applicants,
were referred for a psychiatric examination, the results of which could
constitute grounds for ineligibility.

   New screening standards. In August of 1968, a program for rapid
expansion of the National Airspace System was initiated and a revised
set of trainee-selection standards was adopted. The new selection
program was highly similar to that of the preceding 56 months, but two
new screening standards were implemented as a means of insuring an ade-
quate supply of candidates (Cobb, Young, and Rizzuti, 1976). One of the
new standards allowed applicants with highly specialized ATC experience
(particularly in radar control) to be granted waivers of the aptitude-
screening requirement and also to be appointed to training at pay grade
GS-9 or higher rather than the normally prescribed entry grade of GS-7
or lower. It was reasoned that such personnel would be able to complete
developmental training more rapidly than others and would more quickly
alleviate the shortage of full-performance-level controllers. The second
standard established a score of 210 on the CSC test battery as a common
screening standard for most other applicants. The screening standard of
210 applied to: (1) former military controllers unable to qualify under
the "specialized experience" standard, (2) pilots, navigators, air dis-
patchers, and others who would have confronted a test-score screening
hurdle of 225 if they had applied in earlier years, (3) 4-year college
graduates with records of superior academic achievement, and (4) appli-
cants having master's degrees. A test score of 240 was prescribed for
screening of applicants with no aviation-related experience, most of whom
were college graduates of less than superior academic achievement (Cobb,
Young, and Rizzuti, 1976). The "specialized experience" standard remained
in effect until April 1972. Throughout that time, however, less than one-
fourth of the ATCS selectees entered as GS-9's or higher with waivers of
the aptitude-screening requirement (Cobb and Nelson, 1974).

Post-Academy Attrition

   Only 710 (18.9%) of 3,751 trainees who arrived at the FAA Academy
during November 1968 and the ensuing 17 months for basic ATCS training

86.

entered the FAA as GS-9's or higher on the basis of highly specialized
ATC experience (Cobb, Lay, and Bourdet, 1971); 446 of these enrolled in
the Academy's basic En Route course, and 264 entered the Terminal course.
A study revealed that 16% of the 446 En Route trainees of GS-9 level and
higher failed the En Route course, compared to 18.2% for the 2,526 who
enrolled in the same training course as GS-7's or lower. Only 14% of
the 264 Terminal trainees recruited with waiver of aptitude-screening
failed the Terminal course, compared to a significantly higher elimina-
tion rate of 21.9% for the remaining 515 Terminal students (Cobb and
Nelson, 1974). However, a later CAMI study by Cobb, Mathews, and Nelson
(1972) in which December 1, 1971, served as a common date for determina-
tion of the attrition-retention status of every student who successfully
completed either EnRoute or Terminal basic training at the Academy during
1969, indicated that the trainees selected under the "specialized ex-
perience" standard had slightly higher post-Academy attrition rates than
those who were appointed to training at pay grades of GS-7 and lower.
The difference between the post-Academy (i.e., facility-training) elimina-
tion rates of the two differently selected Terminal subgroups was somewhat
greater than that obtained between the EnRoute subsamples, but neither
difference was statistically significant. Had the results not been con-
founded by aging effects, Cobb and his associates would have concluded
that specialized ATC experience was of little or no value to most trainees
after Academy graduation. However, almost 23% of the higher graded trainees
of the combined EnRoute and Terminal options were 35 years of age or older,
whereas slightly less than 14% of those appointed as GS-7's or lower were
older than 34 (Cobb and Nelson, 1974).

Unpublished research by Cobb, cited by Cobb and Nelson (1974), in-
volving several hundred ATCSs who had successfully completed either En-
Route or Terminal basic training at the Academy in 1969, revealed highly
significant differences between the post-Academy attrition rates (by
December 1, 1971) for trainees aged "35 and older" vs. those "34 and
younger." For Academy graduates at GS-9 level and higher, the facility-
training attrition rates were 42% and 17.5% for the older and younger
subgroups, respectively. About 25% of the ATCSs at GS-7 level and lower,
who were over 35 years old, attrited after returning to their home faci-
lities, whereas the post-Academy elimination rate of the younger ATCSs
having similar pay grades was only 18%. Moreover, several earlier studies
(Cobb 1968a; Cobb, Mathews, and Lay, 1972; Trites, 1964; Trites and Cobb
1964a) had consistently shown chronological age to be inversely related
(at highly significant levels) to scores on numerous aptitude tests,
various indices of Academy training progress, and ratings of journeyman-
level job performance.

As early as 1965, it was the view of some FAA officials that a special
early retirement program was needed for controllers and that the recruit-
ment of ATCS trainees should be restricted to qualified applicants who
were relatively young. . However, such proposed policies ran counter to
the CSC regulations pertaining to all Federal service employees except
those specifically exempted by congressional legislation. Research con-
cerning age-related effects upon ATCS performance was intensified and,

in 1972, the cumulative body of findings prompted congressional legis-
lation authorizing the FAA and CSC to establish a screening standard
with respect to age (age 30) for all ATCS applicants and to develop and
implement a proposed ATCS "Second-Career Program." The congressional
bill, Public Law 92-297, became operational on August 14, 1972. Since
that time, ATCSs receive credit for 1.4 years of Federal service for
each year of active control work; the normally prescribed minimum-age
requirement of 55 does not apply to control personnel; early retirement
is not mandatory but retention as an ATCS requires maintenance of job
proficiency, and, for a time, ATCSs were also offered training for othe
jobs (i.e., "Second-Career Training"; this section of the legistlation
has not been funded by Congress since 1979).

Selection and Training Programs to Aid Minorities and Women

During the 1960's, Government-wide awareness of the social need to
provide opportunities for certain economically and culturally disad-
vantaged groups prompted the FAA to implement a new source of recruiting
for the ATCS occupation. The new project was originally proposed in an
FAA Organizational Bias Seminar, and was committed by the Administrator
to the Secretary of Transportation in 1968. It was termed the "150
Program." The purpose was to provide a predevelopmental program, with
entry into the ATCS occupation at the GS-4 level rather than the previous
GS-5 minimum accession level. The Predevelopmental program was designed
to provide training primarily for persons of underprivileged backgrounds,
as a means of hiring increased numbers of minorities in ATCS positions,
which constitute about half of all FAA positions.

The program involved a training agreement with the CSC, in which
positions were allotted on a 6-month recycling basis. Special 6-month
qualification training was given at the FAA Academy upon the trainee's
entrance on duty. The courses were designed to provide the GS-4 ATCS
with a general background of aviation knowledge prior to entering the
Academy. On October 1, 1969, the CSC approved training agreements
covering ATCSs for a 2-year duration. This allowed for accelerated
promotion from the GS-4 entrance level to GS-5, after completion of 6
months of Academy training (FAA Office of Personnel, Note 2).

Since inception of the "150 Program," evaluation has taken place
in several ways. For the first 21 ATCS trainees who entered the Academy
on February 9, 1970, biographical backgrounds were compared with proposed
course content, and course refinements were introduced. Those classes
were graduated in July 1970. End-of-course critiques completed by trainees
and instructors were evaluated, and further critiques were mailed and com-
pleted at field installations where graduates were subsequently assigned
in February 1971. Several refinements in the program were made as a
result of the initial evaluation process (FAA Office of Personnel, Note 1).

In early 1971 the FAA Office of Personnel conducted an informal evalua-
tion based primarily on discussions with students, faculty members, and
regional personnel staff. Refinements were made in recruiting efforts as
a result of this study. A more formal evaluation of the "150 Program" was

88.

performed by the Office of Personnel in September, 1971. The general conclusion of the study was that the program "provided an additional rung in the ladder" for a high percentage of minorities who otherwise would not have become agency employees. A later study by the same office verified this general conclusion (FAA Office of Personnel, Note 2).

Boone (1978) performed a path analytic study of the "150 Program" to determine whether the program had a direct impact on the trainee's ability to achieve success in Academy training. The two previous informal studies had shown that the "150 Program" resulted in a higher percentage of women and minorities in ATC work; however, no explicit evaluation had been performed to demonstrate that the increase in the number of women and minorities resulted from the instruction received in the program rather than from a mere increase in the number of women and minorities recruited and hired through the "150 Program." The participants in the study consisted of all persons who came through the Predevelopmental program during calendar years 1974 through 1976. The effect of the program on Academy success was reviewed for nonminority men and women and for minority men (there were too few minority women to form a basis for analysis). The general conclusion of the study was that, overall, the Predevelopmental program aided the disadvantaged to achieve success in the FAA Academy. This general effect, however, was not true for the subgroups. The path models indicated that non-minority men and women (especially women) were aided by the programs. However, minority men (predominantly Blacks) were not; rather, the recruitment and selection testing procedures could have produced this differential effect.

Research Related to the Uniform Guidelines on Employee Selection

The recent development of a set of uniform guidelines on employee selection procedures, rooted in the 1964 Civil Rights Act Title VII, as amended by Congress in 1972, prompted an additional research focus at CAMI beginning in 1976. This coincided with the start of a new pass/fail ATCS training program at the Academy and an accelerated pace of FAA-CSC activity to develop a new ATCS selection battery.

By 1976 two separate sets of guidelines were in existence. One set was developed by The Equal Employment Opportunity Commission (EEOC) and the other by the Department of Justice and the CSC. In 1977, the Carter administration authorized an effort to unify the two guidelines and in 1978, the EEOC, CSC, Department of Labor, and Department of Justice adopted the Uniform Guidelines on Employee Selection Procedures (1978).

The Uniform Guidelines require that "employer policies or practices which have an adverse impact on employee opportunities of any race, sex, or ethnic group are illegal... unless justified by business necessity." The guidelines go further in defining adverse impact as a selection policy that results in a selection proportion for any race, sex, or ethnic group that is less than 80% of any other race, sex, or ethnic group. Business necessity is defined as a properly performed validation study. The validation

89.

procedures must involve a statistical study; expert or professional opinion is not acceptable in lieu of a proper statistical study.

The 1978 guidelines answered several important questions about discriminatory selection practices. However, they raised other questions specifically related to ATCS selection procedures that would have to be taken into account in efforts to devise a revised battery of ATC aptitude tests. A pertinent feature of these efforts was the fact that data on new tests were based for the most part on ATCS trainees, a population that had already passed screening hurdles, thus restricting the range of scores in test samples. Boone and Lewis (1978) developed a new procedure to make maximum use of available information in the adjustment of validity coefficients for restriction of range due to selection. This new procedure provided a more accurate method to correct spuriously low correlations between selection and measures of ATCS job performance for samples of persons already selected for ATCS training. The procedure was compared to the well-known procedures developed by Gulliksen and Thorndike and gave more accurate corrections, especially in cases in which restriction was extreme.

To assess the impact of the requirement of the Uniform Guidelines, of test fairness to racial, ethnic, and sex groups, Lewis (1979) used Monte Carlo data with characteristics simila· to those of current ATCS selection samples to compare three well-known models of fairness; (1) Thorndike's Constant Ratio Model, (2) Darlington's Conditional Probability Model, and (3) Einhorn and Bass' Equal Probability Model. The models were compared for robustness in relation to sample size differences, different predictor and criterion correlations, and different selection and success ratios. Essentially, the study demonstrated that the models were equally fair under certain specified conditions and the application of the models depended on the goals and aims of the agency involved. Lewis emphasized the modification of minority recruitment practices as an effective means of complying with the Uniform Guidelines without necessitating the development of new selection devices.

In another study, the effect of recruitment procedures on correction of validity coefficients for restriction of range was examined by Boone and Lewis (1980). This study demonstrated that different recruitment procedures can result in widely varying (unrestricted) variances in applicant groups. Highly specific restriction results in greater homogeniety in applicant groups, which reduces variances on selection tests and thus reduces the magnitude of validity coefficients. Recommendations were offered in the study to help minimize this problem. Boone (1979b), in response to an expressed need from the FAA Office of Aviation Medicine, delineated a mathematical procedure for eliminating outlier data in distributions of experimental aptitude test scores which have a small probability of belonging to the remaining group of data under study. Basically, the process assumes a multivariate normal distribution and uses the Maholanobis D distance function to determine the probability that a particular data vector belongs to the remaining group of data. Examples applying the procedure to experimental test batteries for ATCS selection were given. The products of these efforts are currently being used in continuing validation research at CAMI.

DEVELOPMENT OF NEW SELECTION TESTS:
CONTRIBUTIONS TO THE 1981 OPM SELECTION BATTERY

Research Toward a New Selection Battery

Recent efforts to update and improve ATCS selection procedures were rooted in part in recommendations by a task force commissioned by the FAA in December 1974 to review the agency's selection policies. These were recommendations to concentrate further research and development in the following areas: (1) the testing and screening of applicants for ATC work, (2) the CSC rating guide used to grant additional points for certain types of related prior experience, and (3) the evaluation of further recruitment and testing practices for cultural bias against women and racial minorities. As a result of that review, major studies were begun almost immediately and extensive new research and development activities were undertaken at CAMI and in the FAA Office of Aviation Medicine, (described in Part IV Chapters 15 through 23). These efforts contributed to the new test battery adopted in the Fall of 1981.

The EPA studies. The first of the new studies was carried out under FAA contract by Colmen and his associates at Education and Public Affairs (EPA), a private company, to determine the potential of an experimental test battery to predict ATCS success, defined by a composite of supervisory assessment and career progression measures: This study is summarized in Chapter 18, Part II, and was an extension and expansion of an earlier EPA contract study in 1972 (see Chapter 18, Part I).

The EPA work provided data-based predictions based on samples of currently employed ATCS's. A later study by Boone (1979, see Chapter 18, Part III) investigated the prediction of Academy laboratory grades on a sample of applicants, using most of the same tests.

The Boone study. Boone's study undertook further evaluation of the relative value of the tests that had demonstrated potential in the EPA studies. It was enabled by the Civil Service Commission, which in 1977 administered two of the experimental instruments, the Multiplex Controller Aptitude Test (MCAT, see Chapter 15) and the Directional Headings Test (DHT), to approximately 7,000 ATCS applicants in conjunction with the regular CSC ATC test battery. These data were used by Boone (1980) along with data collected by CAMI on new recruits to correct for restriction in range in predictor data (Boone 1980). The corrections were made using a previously developed new procedure (Boone and Lewis, 1978; 1980). The data were based on trainees selected for ATCS work beginning May 1976 through April 1978. During their first day at the Academy, new trainees were tested by Boone and Lewis with experimental test batteries which included a Biographical Questionnaire, a Dial Reading Test (DRT), the MCAT, and the DHT. Only ATCSs who had a complete data set were included in the final sample of 1,828 trainees.

The DRT used in the Boone study was Part I from the U.S. Air Force's Dial and Table Reading Test (Guilford and Lacey, 1947); The examinee is

91.
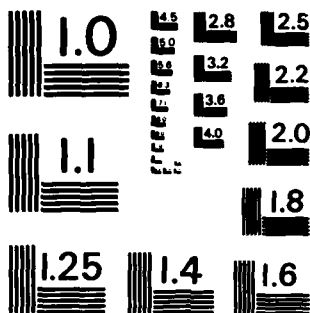
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

presented with seven dials for each set of six questions and is required to read the correct value on the appropriate dial in order to select the right answer from among five given alternatives. The MCAT (Dailey and Pickrel, 1977, see Chapter 15) presents pictures of simulated air traffic crossing a controller's display, and a table of altitudes, speeds, routes, and identifiers. The test is speeded and reflects changes in aircraft position in successive items. A primary task is to predict violations of aircraft standards while other questions measure aptitudes such as table reading, spatial visualization, and arithmetic reasoning. The DHT (Cobb and Mathews, 1973) is a highly speeded test of perceptual-discrimination and coding skills developed at CAMI. Items on the DHT present one to three pieces of information reflecting the cardinal points on a compass. The examinee must very quickly determine the compass direction (or its exact opposite) represented by the pieces of information and determine if the bits of information agree or are in conflict. Promising results obtained by Cobb with this test had been confirmed by Milne and Colman (1972) in a study of journeymen controllers.

Scores on the DRT, MCAT, and DHT and on the tests comprising the 1974 CSC battery were correlated by Boone (1980 see also Chapter 21) with averaged laboratory scores from Academy training. The tests that made a significant contribution in predicting Academy scores were then used to form composite scores, and six regression models were explored, representing combinations of the CSC tests and the new (experimental) tests. Two of the combinations demonstrated a statistically significant increase in the multiple correlation over the old test battery. Of the two models, one involved use of the MCAT with two current CSC tests (CSC 24, Computations; CSC 157, Abstract Reasoning), the other involved no current CSC tests but included the MCAT, DHT, and DRT.

Based on the analyses by Boone at CAMI, by EPA, and on later studies (in 1981) by CAMI and OPM regarding fairness of the tests (see Chapter 22), a new selection battery was formed consisting of two versions of the MCAT and one version of the Abstract Reasoning Test from the earlier battery. The new test battery was officially adopted for use in October 1981. By coincidence, this occurred just in time to meet the demand for thousands of new ATCSs following the ATCs strike and the subsequent firing of approximately 12,000 controllers (see Chapter 5). Data from the new selection tests are being maintained and monitored as a part of the CAMI data base.

Current Status of Selection Standards

The ATCS selection procedures remained essentially unchanged from August 1968 until April 1973. At that time, two important procedural changes in the selection process were made: (1) selection at the GS-9 level on the basis of "specialized experience" ceased (April 2, 1973) and mandatory aptitude-screening was reinstated for all applicants, and (2) the selection procedure was modified to limit eligibility of applicants to those 30 years of age or younger. With the exception of these two aspects, the 1973 selection program was highly similar to that of 1968.

From 1973 until October 1981, the selection process remained essentially unchanged. However, the period from 1973 until August 1981 (when the ATCS strike occurred) saw a drastic decline in recruiting demand. In fact any candidate during this era had little chance for an appointment unless his or her overall eligibility rating far exceeded that which would have warranted selection during the 1968-73 period. In addition, the Academy training program was modified in 1976 to include new pass/fail criteria based on a curriculum that assumed zero ATC knowledge on the part of trainees. CAMI was assigned the task of maintaining the longitudinal data base for all trainees beginning with the revised Academy program. That computerized base includes CSC (now OPM) selection test scores, biographical information, all Academy training scores, and subsequent field performance data on Academy graduates.

In the final quarter of 1981, subsequent to the ATCS strike and the firing of striking controllers, the new selection battery was made operational. Simultaneously, use of the Occupational Knowledge Test (OKT, see Chapter 16) was initiated. OKT scores are now used to add 0, 3, 5, 10, or 15 points to the aptitude test composite, after the applicant establishes eligibility by scoring 70 or better on the new battery and by meeting the experience and/or education requirements. OKT points replace the prior method of granting extra points on the basis of related experience. Ranking on the register is achieved on the basis of this total score, plus any applicable veterans preference points. A sharply increased recruiting schedule was implemented in October 1981 as part of the "Air Traffic Recovery Program." This schedule resulted in the selection of over 500 former military ATCSs at the GS-9 level (based on similar criteria used to select GS-9s during the 1968-73 period) and drew 125,000 applications for ATCS positions. By January 1982, approximately 26,000 applicants were certified as eligible on the ATCS OPM registers. An estimated 8,000 new air traffic controllers were planned to be added to the workforce by the end of 1983.

SIGNIFICANT AREAS INVESTIGATED AT CAMI

Attrition During and Subsequent to Training

It had been ascertained in CAMI's primary validation study with the 893 pre-1964 ATCS trainees that Academy-basic-training attrition rates were: (1) 30.3% for the total group, (2) only 16.4% for the 317 who attained composite raw scores of 210 and higher on the CSC test battery, and (3) 38% for the group of 576 with scores of 209 and lower (see Figure 3). Adoption of the CSC battery for operational screening purposes resulted in a reduction of the attrition rate at the Academy from approximately 30% to about 22% for the 2,822 men who met the aptitude screening requirement and who entered the Academy during November 1968 through March 1970 (Cobb and Mathews, 1973).

Cobb, Mathews, and Nelson (1972) took a detailed look at the variations in characteristics of the ATCS trainees and their attrition rates for the periods 1960-63 and 1968-70. This study involved comparisons of Academy basic training elimination rates and post-Academy attrition and retention

CSC Score:	129 & <|	130 - 149	150 - 169	170 - 189	190 - 209	210 - 229	230 - 249	250 & >	N Total
N Pass:	13	33	52	124	135	126	87	52	622
N Fail:	38	46	47	51	37	25	23	4	271
N Total:	51	79	99	175	177	151	110	56	893
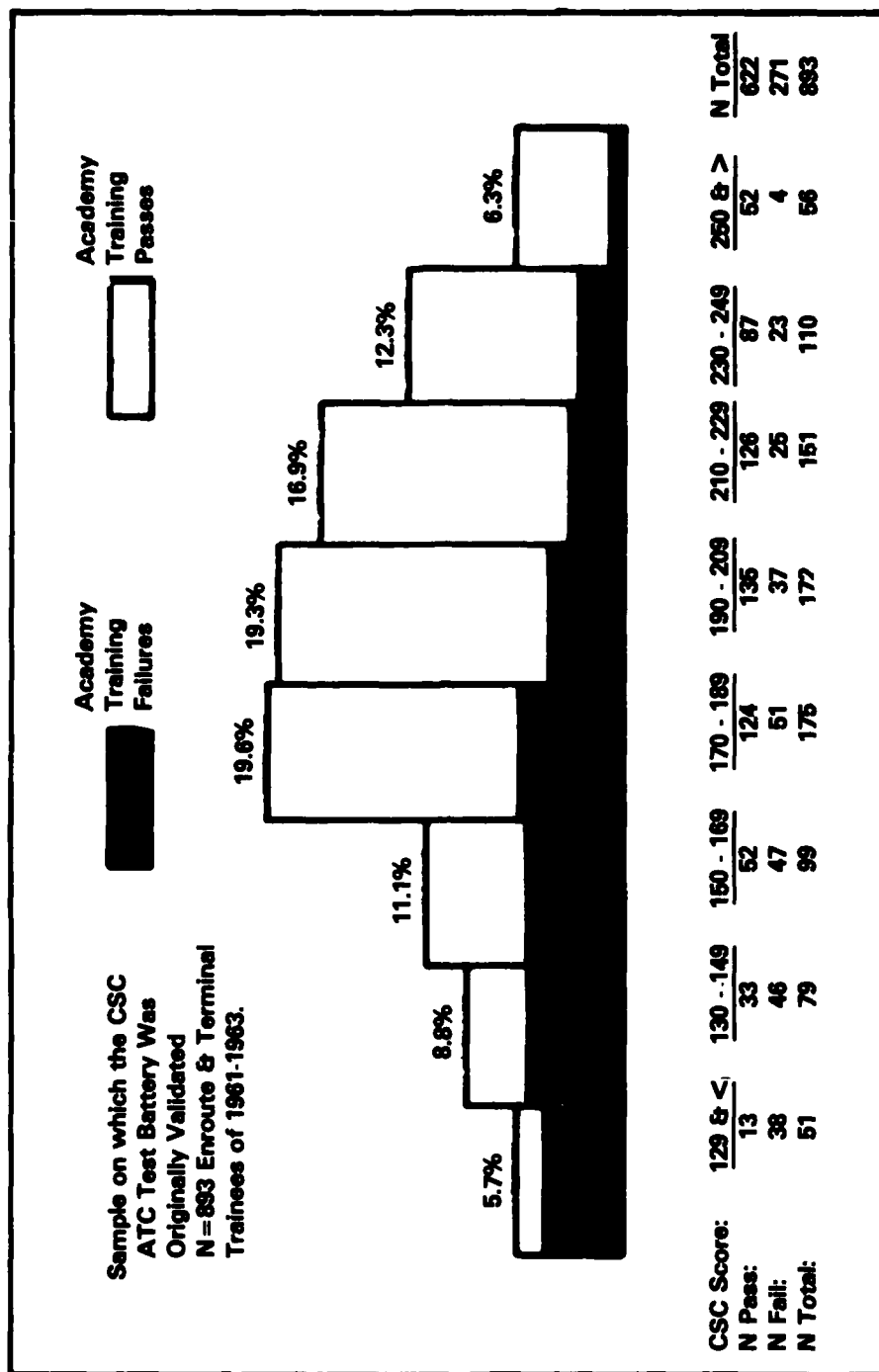
Figure 3. Distribution of CSC ATC Aptitude Screening Test Scores for the 893 Experimentally Examined Entrants Into Academy ATCS Training During 1961-63 (from Cobb and Mathews, 1972).

94.

rates of personnel who were recruited during these widely separated time periods for each of three different types of ATCS training. Data were obtained for a total of 6,367 former trainees. Exactly 2,000 of these trainees had entered the Academy during September 1960 through August 1963, before the CSC battery became operational in the screening of most applicants. The 2,000 included 733 Terminal trainees, 1,008 EnRoute trainees, and 259 FSS personnel. The remaining 4,367 trainees, the vast majority of whom were required to qualify on the CSC test, had entered the Academy during October 1968 through March 1970; of the 4,367, 935 were Terminal trainees, 3,159 were EnRoute personnel, and 273 were FSS personnel.

Percentages reflecting the Academy elimination rates for the earlier versus the later time periods, respectively, were: 20.9% and 19.3% for the Terminal personnel, 32.0% and 17.9% for the EnRoute trainees (this difference was statistically significant), and 18.5% and 12.8% for the FSS personnel. The mean Academy elimination rate of 26.2% for the 2,000 pre-1964 trainees was significantly higher than the 17.0% rate obtained for the 4,367 more recent recruits (Cobb, Mathews, and Nelson, 1972). Moreover, the 17.9% rate was only slightly higher than a rate of 16.4% that Cobb had projected on the basis of results obtained in an earlier study of the 893 pre-1964 trainees examined experimentally with the CSC battery before it was adopted for operational use.

Followup procedures were employed in which Academy graduates who were still in ATC work on December 1, 1971 were designated as "retentions," while those eliminated after completion of Academy training were designated as "post-Academy attritions." The post-Academy attrition rates for the Terminal, EnRoute and FSS entrants of 1960-63 were 16.0%, 22.8%, and 18.1% respectively, whereas the corresponding rates for the recruits of 1968-70 were 10.1%, 20.3%, and 5.9%. The EnRoute option was the only one for which the difference between the rates was not statistically significant.

Analyses pertaining to subgroups, each comprising one to four former incoming classes of the Academy's Terminal, EnRoute or FSS training courses, yielded significant inverse relationships between the Academy elimination rates and post-Academy attrition rates (see Figure 4). Such findings were remarkably consistent for personnel who entered each type of training during either of the two time periods. These findings also suggested the possibility that the basic training courses could have been utilized more effectively to prevent the advancement of some borderline and un-qualified recruits to facility training from which they were eventually eliminated.

An analysis for 645 Academy Terminal graduates and 2,162 EnRoute graduates between January 1969 through March 1970 revealed a post-Academy retention rate for Terminal trainees of 87.8% and for EnRoute a signifi-cantly lower rate of 75.7%. A series of analyses in which no distinction was made with respect to trainees' pay grades or qualifications at entry into training indicated that the retention rates varied appreciably from facility to facility. However, the retention rates did not appear to

	Oct 1968	Nov-Dec 1968	Jan-Feb 1969	Mar-Apr 1969	May-Jun 1969	Jul-Aug 1969	Sep-Oct 1969	Nov-Dec 1969	Jan-Feb 1970	Mar 1970	Total
Retentions (Dec. 71)											
N Cases	120	326	102	200	538	356	166	208	588	233	2,837
Percent of Group	79.5	74.6	65.4	69.4	60.0	63.7	57.6	61.0	65.1	67.0	65.0
Post-Academy Attritions											
N Cases	29	107	42	54	171	78	45	52	121	51	750
Percent of Group	19.2	24.5	26.9	18.8	19.1	14.0	15.6	15.2	13.4	14.7	17.2
Academy Attritions											
N Cases	2	4	12	34	187	125	77	81	194	64	780
Percent of Group	1.3	0.9	7.7	11.8	20.9	22.4	26.7	23.8	21.5	18.4	17.9
Total N of Group	151	437	156	288	896	559	288	341	903	348	4,367

Figure 4. Percentages of Entrants Into the Academy's EnRoute, Terminal, and FSS Courses During October 1968 Through March 1970 Who Were Academy Attritions, Facility-Training Attritions, or Were Still in FAA ATC Work in December 1976 (from Cobb, Mathews, and Nelson, 1972).

be related, positively or negatively, to the traffic-density levels of the facilities. For example, the mean retention rate of assignees to the 12 Level-IV (or top-ranked) Terminal facilities was 89.2%, whereas the retention rates of trainees averaged 87.9% at all Level-I and Level-II facilities and 87.0% at the Level-III installations. Although significantly lower than those obtained for Terminal personnel, the mean retention rates of EnRoute personnel at the 13 Level-II Centers (75.4%) and 14 lower ranked Centers (75.8%) were highly comparable.

Cobb, Mathews, and Nelson (1972) also found that personnel who qualified for entry into training at the GS-9 level and higher on the basis of pre-FAA specialized experience had significantly higher Academy graduation rates than the conventionally recruited trainees. However, at the Level-IV and Level-III Terminal facilities, post-Academy attrition rates of the higher rated trainees averaged about twice that of the less experienced trainees, also, their mean attrition rate at the 13 top-ranked Centers was higher, though not significantly so, than that of other trainees. It was only at the Level-I EnRoute Centers and Terminal facilities of the two lowest levels that the mean attrition rates reflected favorably on these more experienced trainees--and then only by one percentage point.

Although other types of occupations that make use of initial screening tests and formal centralized training are not directly comparable to air traffic control work (nor to each other), their attrition rates were examined to provide at least some perspective from which ATC Academy attritions of 1968-70 might be viewed. For example, published reports available at that time (circa 1970) about several occupational groups that were carefully prescreened prior to entry-into-training revealed training-attrition rates ranging from 22% to 43%. Among these were attrition rates for Peace Corps volunteers (Gordon, 1967); nursing students in their first year of training (Katzell, 1968); nursing turnover in a teaching hospital (Saleh, Lee, and Brien, 1965); Army officer candidates during 23 weeks of training (Peterson and Lippitt, 1968); and naval aviation students (Bale and Ambler, 1971).

Since ATCSs make up about half of the FAA work force, some comparisons of their annual attrition rates with rates of non-ATCS personnel provided another occupational perspective. The same caveat noted above with regard to comparison of different types of occupations (as well as different GS-grade levels and other characteristics), applies to these comparisons. Cobb, Mathews, and Nelson (1972) calculated the average attrition rate for all FAA employees not in the air traffic control occupational specialty and it was approximately 13.5%, 9.3%, and 9.1% for calendar years 1969, 1970, and 1971, respectively (averaging 10.6%). The rates during the same time periods for all ATCSs were 5.7%, 7.8%, and 5.0% (averaging 6.3%). If the assumption were made that those who attrited from the FAA Academy would have been later ATCS attritions anyway, then, following selection and Academy screening, the attrition rates of ATCSs would have dropped to approximately 3.3%, 5.6%, and 4.9%, respectively, for 1969, 1970, and 1971 (a 3-year attrition rate averaging 4.6%).

These 3-year attrition rates were also examined by means of longitudinal comparisons between ATCSs and noncontroller FAA personnel hired from October 1968 through March 1969. During that period, 949 ATCSs (including many who entered the Academy during April or later) were hired at the GS-5 through GS-9 levels; a total of 217 non-ATC employees were also hired at those same levels. The overall attrition rate (as of December 1971) for the 217 noncontrollers was 32.7%; for the 949 ATCSs, the rate was 24.6% (including Academy failures).

Age Limit for Applicants

Perhaps the most important of all the CAMI studies on ATCS selection were those demonstrating the effects of age on training and job performance. Training records for groups of ATCS personnel recruited since 1960 and a number of unpublished earlier CAMI studies consistently revealed a definite relationship between chronological age and probability of attrition.

To enter ATCS training at the time CAMI began its research program, an individual had to be at least 21 years old, but there was no upper age limit. It had long been felt by training and supervisory personnel even in 1960 that a marked negative relationship existed between age at entry into training and subsequent performance.

In CAMI's first assessment of this relationship, the data in Figure 5 were compiled by Trites and Cobb (1964a) from several different samples of ATCS trainees. In the figure, "Fails" were those who did not complete training, "Separated" were individuals who completed training but were no longer employed by the FAA, "Marginals" were individuals who were still with the FAA but about whom a job supervisor had some reservations concerning adequacy of job performance or potential; and "Satisfactory" individuals were those who were working satisfactorily in their ATCS specialties. The negative relationship between age at entry into training and job performance is obvious. The older the trainee, the less likely he or she was to complete training or to remain with the FAA. In an unpublished 6-year followup study on 688 trainees (who comprised the successive classes of the Academy's basic training courses during August 1961 through March 1963), Cobb found that (1) 72% of 118 men age 36 and over failed their initial training course whereas only 33% of 570 younger students failed, and (2) older trainees who did pass the Academy also tended to experience greater difficulty than their younger colleagues in subsequent phases of on-the-job training. In a later published study that included these ATCSs and later trainees, entry age proved to be inversely related to both the aptitude test scores and criterion measures (Cobb, 1968a). The consensus that older controllers generally performed their duties less effectively than younger ATCSs was supported by several studies (Cobb, 1968a; Trites and Cobb, 1964a, b), all of which indicated that performance, as evaluated by supervisors or peers, was apt to decline after age 40 without regard to tenure or experience in ATC work (Cobb, Lay, and Bourdet, 1971).

These and other CAMI investigations, dating back to 1961, also indicated that the training attrition rates of groups of trainees over age 30
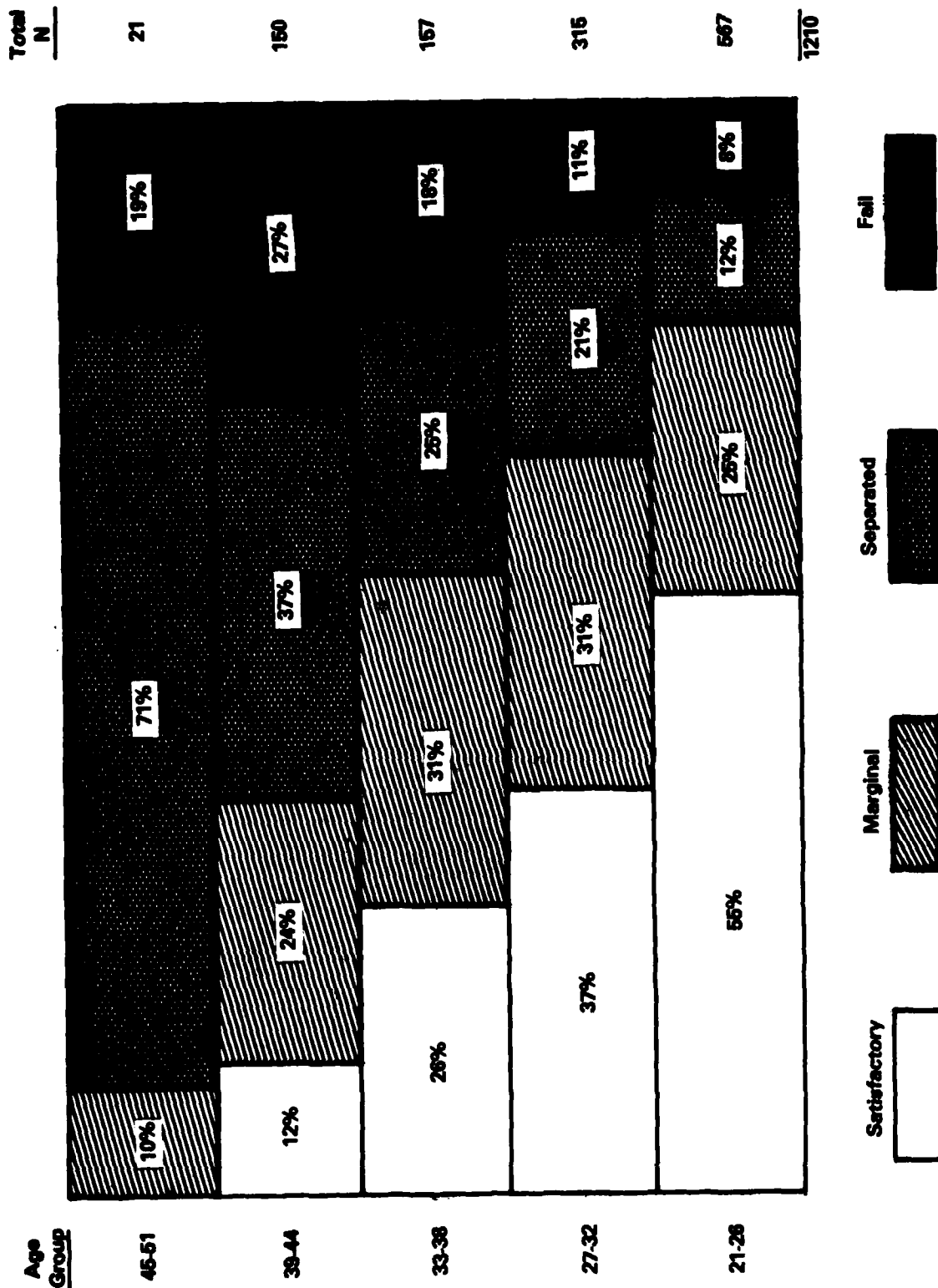
Figure 5. Percentages of ATCSs by Age Groups in Four Training and Job Performance Categories: Satisfactory, Marginal, Separated, and Failed (from Trites and Cobb, 1964).

were generally two to three times higher than those of the younger trainees (see Table 1). In studies by Cobb, Nelson, and Mathews (1974) in which experimental ratings of job performance were collected on journeyman-level ATCSs from both their supervisors and their peers, the mean performance ratings of controllers within every age category beyond 40 were significantly lower than those of the younger subgroups (see Figure 6). Such findings and other unpublished analyses that Cobb provided at the request of the FAA and the CSC played a decisive role in enabling the FAA to obtain Congressional enactment of legislation in 1972 permitting the establishment of an optional early retirement program for controllers and also the imposition of an upper age limit of 30 in the recruitment of controller trainees. (Note: Neither the maximum age limit of 30 nor the early retirement program applies to FSS personnel.)

Aviation-Related Experience as a Selection Factor

Throughout the history of the FAA and CAA, ATCS selection programs have included standards predicated on the philosophy that almost any type of aviation-related experience should be of value for prediction of success in ATCS training and work. Inasmuch as previous experience in air traffic control (usually acquired in military service) has always been considered of paramount importance, standards have invariable prescribed that it be heavily weighted, directly or indirectly, as a selection factor. Other types of aviation experience traditionally regarded as important, but generally weighted more moderately than prior ATC work, included experience (military or civilian) as aircraft pilot, navigator, communications expert, radar surveillance specialist, and flight dispatcher. Prior to implementation of mandatory aptitude-test screening procedures in 1964 (and exclusive of brief trial periods for procedures that resulted in selection of relatively few trainees), the eligibility ratings of medically qualified ATCS applicants were determined primarily on the basis of assessment of aviation-related experience and education.

Briefly stated, selection programs have always been formulated to maximize the recruitment of controller trainees who, in addition to other qualifications, possessed previous ATC experience. The appropriateness of this policy had been confirmed repeatedly by CAMI followup studies of personnel who entered ATCS training during the decade ending in 1970 (Cobb, Young, and Rizzuti, 1976). Unfortunately, however, the pool of former military controllers has diminished over the years, and the FAA has necessarily recruited increasingly greater proportions of ATCS trainees with other aviation backgrounds and also from those having no aviation experience of any type--but who qualified on the basis of aptitude-test measures and of education.

Cobb and Nelson (1974) reviewed several unpublished studies in which biographical data were collected and analyzed for large samples of ATCSs recruited from the 1960's through the early 1970's, and concluded that in general, 40% or more were former military controllers and that 40% to 45% of the remaining selectees held aircraft-pilot ratings. The results of several early CAMI studies cited above suggested that various types of

100.

Table 1

Comparison of Attrition (Attr) and Retention (Ret) Rates by CSC-ATC Aptitude Screening Test Score, Type of Pre-FAA Experience, and Dichotomized Age Grouping for 2,349 En Route and Terminal ATCSs Who Entered the Academy (Acad During 1968. P-A Refers to Post-Academy Attrition (from Cobb & Nelson, 1974)

CSC ATC Aptitude Test Score

Age	Pre-FAA Experience Category	Unknown				209 & Lower				210 & Higher				All Aptitude Levels			
		N Tot	Acad Attr Rate %	P-A Attr Rate %	Ret Jan 1973 %	N Tot	Acad Attr Rate %	P-A Attr Rate %	Ret Jan 1973 %	N Tot	Acad Attr Rate %	P-A Attr Rate %	Ret Jan 1973 %	N Tot	Acad Attr Rate %	P-A Attr Rate %	Ret Jan 1973 %
81 & Older	Non-ATC Pilot	71	57.7	21.1	21.1	1	100.0	0.0	0.0	115	27.8	33.0	39.1	187	39.6	28.3	32.1
	ATC & Pilot	16	25.0	18.8	56.3	3	33.3	66.7	0.0	9	0.0	22.2	77.8	28	17.9	25.0	57.1
	Non-Pilot ATC	125	35.2	26.4	38.4	40	27.5	25.0	47.5	48	10.4	25.0	64.6	213	28.2	25.8	46.0
	Other	84	64.3	10.7	25.0	5	60.0	20.0	20.0	148	27.0	26.4	46.6	237	40.9	20.7	38.4
	Total	296	48.3	20.3	31.4	49	32.7	26.5	40.8	320	24.1	28.4	47.5	666	35.5	24.7	39.8
80 & Younger	Non-ATC Pilot	97	35.1	21.7	43.3	0	0.0	0.0	0.0	246	9.0	20.4	70.6	342	16.4	20.8	62.9
	ATC & Pilot	29	13.8	20.7	65.5	11	0.0	0.0	100.0	24	0.0	29.2	70.8	64	6.3	20.3	73.4
	Non-Pilot ATC	277	17.3	15.5	67.1	88	6.8	20.5	72.7	350	3.1	19.4	77.4	715	9.1	18.0	72.9
	Other	165	54.6	15.8	29.7	3	0.0	0.0	100.0	395	9.9	20.3	69.9	563	22.9	18.8	68.3
	Total	568	31.0	16.9	52.1	102	5.9	17.7	76.5	1014	7.1	20.2	72.7	1684	15.1	18.9	66.0

101.
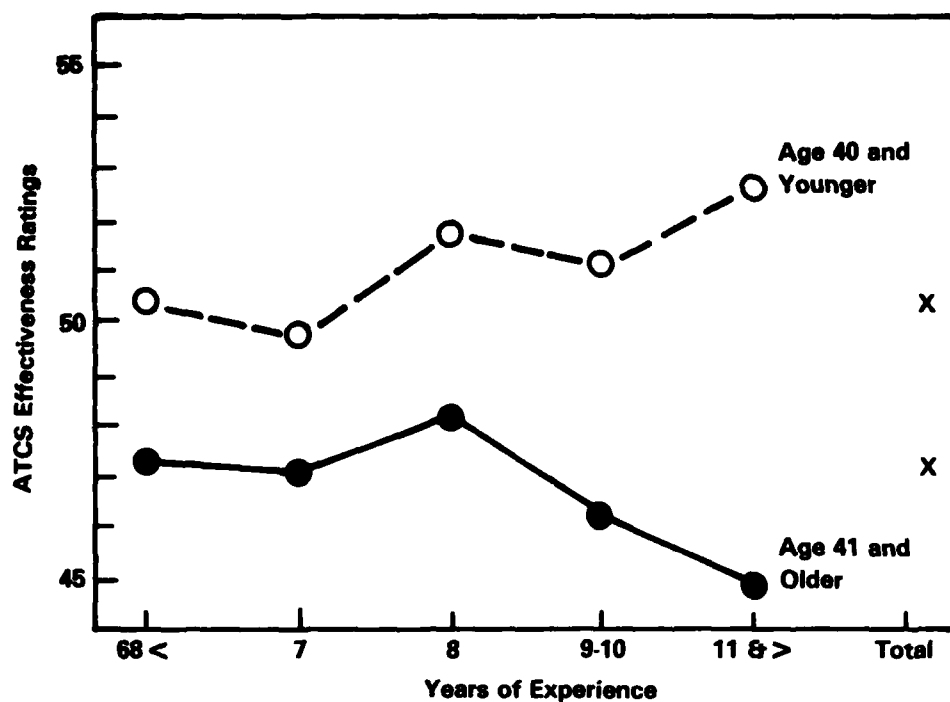
Figure 6. Means of Ratings Combined from Supervisors and Peers for "Older" vs. "Youn
Radar ATCSs of Different Experience Groups (Adapted from Cobb, 1967).

102.

pre-FAA ATC experience were beneficial to the ATCSs primarily during the basic training phase only, whereas all other experience, including aircraft-pilot experience, appeared to be of questionable value at any stage of the training. Yet, little or no information had been gleaned through the 1960's to indicate whether the pilot-experience standards should be abolished, drastically revised, or modified only slightly.

Under the standards, a total of five points was (until the adoption of the new battery in 1981) credited toward the overall eligibility rating of each pilot-rated applicant having 350 or more hours of logged flight time. A cursory review of biographical-questionnaire response data for several hundred pilot-rated ATCS trainees recruited during 1969 revealed that about half of them possessed no more than a private pilot license and 350 to 500 hours of logged flying time, and less than 30% had 1,000 hours or more (Cobb and Nelson, 1974).

Increased skepticism regarding the validity of pilot experience for ATCS selection followed publication of the Corson Report (1970) of the Air Traffic Controller Career Committee in January 1970. This report stated that no evidence could be found indicating any type of pre-FAA experience other than ATC work to be useful for prediction of FAA ATCS training progress or subsequent job performance. The committee recommended "elimination of credit for pilot experience" in the selection process. The same recommendation was made in a contract study report to the FAA by Colmen (1970). However, neither of these reports cited any studies other than those by CAMI, as a basis for their conclusions and recommendations concerning pilot experience, and the studies cited did not necessarily imply that all such experience should be completely disregarded. None of the studies had focused directly on the issue of pilot experience; none included determination of the attrition-rate probabilities for ATCSs relative to flying time or types of ratings held (e.g., private license, commercial license, instrument rating, air transport rating, etc.); and the interaction effects of age, aptitude, and education on the validities of pilot experience and other types of experience had never been assessed and compared. In order to address the issue a comprehensive longitudinal study was undertaken by Cobb and Nelson (1974).

That study was based on data for 4,092 ATCS trainees and examined the validities of various types of aviation-related experience, separately and in combination, for prediction of success in FAA ATC work. Success was defined as retention status within the ATC system several years after entry into training. Of the 4,092 ATCSs, 1,740 entered Academy basic training during September 1960 through August 1963, before the CSC ATC battery became operational in the screening of most applicants. The remaining 2,352 ATCSs, the majority of whom were selected from among aptitude-screened applicants, entered Academy training during 1969. Both groups entered prior to establishment of the current age eligibility standard of age 30. The results clearly demonstrated that success in FAA ATC work was far more contingent on entry age than on type of aviation-related experience, level of aptitude, or level of education. The findings led Cobb and Nelson to

103.

suggest that ATCS applicants who meet the existing age and aptitude screening standards should not be awarded credit points toward their eligibility ratings for any type of experience other than ATC work (see Figure 7), and that even the latter should be assessed and weighted conservatively in the selection process, particularly with respect to military control experience that involved no instrument flight rule (IFR) operations. Cobb and Nelson noted parenthetically that if such a procedural change were followed, an indirect result would likely be a relative improvement in the competitive ranking of women and minority candidates who, for various sociocultural reasons, probably do not obtain the types of pre-FAA experience for which credit is given in selecting ATCS candidates.

Candidates with only pilot experience when they entered ATCS training during either of the two widely separated time periods had unusually low retention rates, even lower than those of groups with no aviation-related experience of any type. Speculation that many of the pilot-rated ATCSs might have attrited voluntarily in order to take jobs more compatible with their flying interests or at higher pay, prompted a search of the standard personnel records of all FAA employees; only 14 of the 254 attrited pilot-rated ATCSs of 1969 were still in the FAA (in non-ATC jobs) at the beginning of 1973. Of the 14, six were aviation safety officers (five of them were GS-12's); one was a GS-12 in an educational training program, and the other seven were in various jobs (e.g., electronics technicians, flight data aids, wage board employees, clerks, etc.) at lower pay grades. Unfortunately, data were not available for a similar followup study of the remaining 240 attrited pilots who had left the FAA.

ATC experience that involved only visual flight rule (VFR) operations proved to be considerably less valid than ATC-IFR experience for prediction of success in either the EnRoute or Terminal option. Cobb and Nelson therefore concluded that, when possible, all EnRoute and Terminal trainees should be selected from among the best qualified of the aptitude-screened candidates younger than 31 who possessed pre-FAA IFR control experience. However, this recommendation was counter to the CSC's traditional policy of evaluating virtually all types of preemployment experience in the selection of personnel for almost any occupational specialty within the Federal service. Cobb, Young, and Rizzuti (1976) expressed the view that the FAA should press vigorously for changes such that aviation experience other than ATC would be evaluated very conservatively in the selection process.

In a later study, Lewis (1978a), investigated the feasibility of using scores on the OKT, a test of general ATC information, to determine the quality of an applicant's ATC knowledge as an alternative to the giving of extra points for experience. The OKT had been developed from older test items provided by the FAA Academy with the intent of being "job-knowledge specific" (Mies, Colmen, & Domenech, 1977). It was developed to provide an instrument for screening of applicants with previous ATC experience who would enter at higher grade levels (see Chapter 14). The Lewis study revealed that

Type of Experience for which a Rating Was Held	Academy Attritions / Post-Academy Attritions / Retentions Jan. 1, 1973	N
Unknown	66.7% / 33.3%	3
Pilot Experience Only	29.6% / 21.1% / 49.3%	71
Pilot with Communications	23.8% / 23.8% / 52.4%	458
Pilot with Pre-FAA ATC	18.2% / 9.1% / 72.7%	11
Pilot with Both ATC & Communications	8.6% / 23.5% / 67.9%	81
ATC Only	13.6% / 18.7% / 67.7%	849
ATC with Communications	12.7% / 31.6% / 55.7%	79
Communications Only	28.7% / 17.0% / 54.3%	94
No ATC, Pilot, or Communications Total	28.2% / 19.7% / 52.1%	706
Total Pilots with No ATC	24.6% / 23.4% / 52.0%	529
Total Pilots with ATC	9.8% / 21.7% / 68.5%	92
Total Non-Pilots with ATC	13.5% / 19.8% / 66.7%	928
Total Subjects with No ATC or Pilot	28.4% / 19.3% / 52.3%	803
Total Sample	20.9% / 20.5% / 58.5%	2,352

Figure 7. Attrition and Retention Rates by Pre-FAA Experience Categories for 2,352 Entrants into the Academy's EnRoute and Terminal Basic Training Courses During 1969 (Adapted from Cobb and Nelson, 1974).

105.

the assignment of extra credit subjectively on the basis of verified experience (which was usually difficult to define) resulted in a significantly higher failure rate than by use of the OKT. Consequently, Lewis concluded that use of the OKT to assign extra credit for specialized experience was more fair than use of "verified" experience.

Under the October 1981 selection standards, any ATC experience or any FAA or military rating (i.e., control tower operator, center or FSS ATCS navigator, navigator bombardier, IFR pilot, VFR pilot with 350 hours or more as copilot or higher, or Air Carrier Dispatcher Certificate) may be used in lieu of education reuirements to establish eligibility for ATCS applicants who score 70 or better on the new aptitude battery. After eligibility is established, experience bonus points are added only on the basis of the applicant's score on the OKT.

Education as a Selection Factor

The ATCS selection programs have always included a mandatory minimum educational requirement of a high school diploma or evidence, such as a General Education Development (GED) certificate, of an equivalent educational background. Education beyond the high school level, although not mandatory, has traditionally received significant weight, directly or indirectly, in the overall eligibility rating. Selection programs have invariably included provisions whereby applicants without background in aviation, could substitute college-level education for such experience. Regardless of experience, however, applicants with 4-year college degrees usually had no difficulty in establishing training candidacy. As mentioned earlier, in one period, college graduates with records of superior academic achievement were allowed an aptitude test score standard considerably below that designated for screening of comparably experienced and otherwise equally qualified applicants. A similar policy at one time applied to all college graduates, irrespective of academic records. However, the greatest emphasis on education as a selection factor was during 1971 and 1972 when applicants with 4-year college degrees, at least 1 year of graduate work, and 12 months of specialized ATC experience could be granted waivers of the aptitude-test-screening and be appointed to training at grade GS-9 rather than GS-7 (Cobb, Young, and Rizzuti, 1976).

However, with regard to this policy, numerous early CAMI studies (based on trainees who entered ATCS training in the decade of the 1960's) in which education was a variable along with age, preemployment experience, and aptitude test performance, indicated that the training attrition rates of ATCS personnel tended to increase (rather than decrease) in accordance with the preentry levels of education (Cobb, 1964; Cobb, Mathews, and Lay, 1972; Cobb and Nelson, 1974; Trites and Cobb, 1964a; Trites, Miller, and Cobb, 1965). Several unpublished CAMI studies mentioned by Cobb, Young, and Rizzuti (1976), based on trainees recruited during various time periods, had shown that college graduates generally had significantly higher attrition rates than selectees with either high school diplomas only or 1 year or less of college.

106.

With this background, Cobb, Young, and Rizzuti (1976) undertook a
definitive study, focused specifically on education. This was based on
data for 2,352 ATCS recruits who entered the Academy basic training
phase in 1969 (1,858 EnRoute and 494 Terminal trainees). They found
that all educational variables, both before and after adjustment for
age and other selection factors, were negligibly and inversely related
to success in ATCS training as defined by Academy graduation status and
retention in the ATC system 3 to 4 years following recruitment (see
Figure 8).

Major courses of study listed by some 925 of 1,265 ATCS trainees
who attended college were found to have little potential for prediction
of training outcomes. The overall retention rate as of January 1, 1973,
for 1,265 former college students in the sample was 56.7%. When cate-
gorized on the basis of college majors, only those (N=141) who majored
in social sciences had a retention rate (41.8%) that differed signifi-
cantly from that of the combined categories. Moreover, 53 recruits for
whom major college courses were judged to be more directly related to
aviation than all others combined, had a retention rate (56.6%) almost
identical to that of the entire group.

Sex as a Factor in Performance and Attrition

Based on the FAA's interest and participation in programs to eliminate
sex as a discriminating factor in the selection of new Federal employees,
Cobb, Mathews, and Lay (1972) carried out a comparative study of female
and male ATCS trainees. The study was also prompted by the fact that
women had never represented more than a very small proportion of all
personnel directly involved in air traffic management. Since experience
is heavily weighted in calculations of eligibility ratings, the "best
qualified" by normal standards are usually male veterans with experience
in military ATC work or as pilots (Cobb, 1962, 1964; Trites, 1961; Trites
and Cobb, 1964a, 1964c).

The study compared age, education, pre-FAA experience, aptitudes,
training-course performance measures, and post-Academy attrition rates
of the 83 women who entered basic ATC training at the FAA Academy during
the 17-month period between November 20, 1968 through March 27, 1970 with
those of various samples of the 3,760 males who entered training during
the same period. There were no significant differences between the means
of the female and male trainees with respect to age and educational level.
On performance on 36 different aptitude tests, only four mean differences,
all of which favored the women, proved statistically significant. Only
45.8% of the 83 women had pre-FAA ATC-related experience, compared to
63.9% for a sample of 798 men; the difference was statistically signi-
ficant. The means of training course grade averages for the two groups
differed by only three-tenths of one point, and there was no significant
difference between the Academy attrition rate of 20.5% for the women and
23.2% for the 798 men. However, the groups differed markedly with respect
to post-Academy attrition rates; as of December 1971, 48.5% of the 66
women who completed Academy basic training but only 22.5% of the 613 men
(within the sample of 798) who graduated from the Academy attrited sub-
sequently.
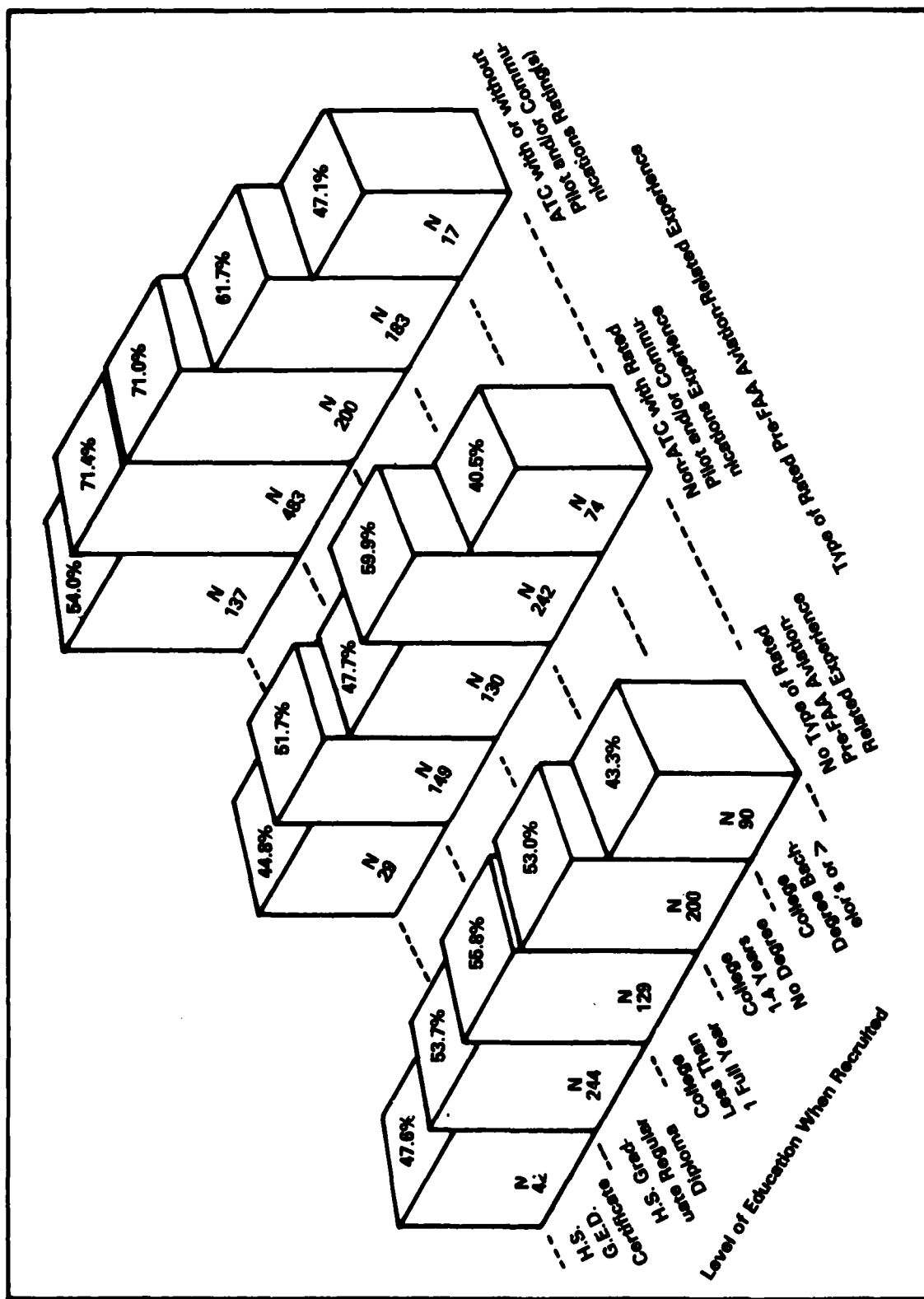
Figure 8. Retention Rates by Level of Education and Type of Rated Pre-FAA Aviation-Related Experience for 2,349 Selectees Who Entered Academy Basic Training in En-Route or Terminal ATC Procedures During 1968. Rates Reflect Proportions of ATCSs In Each Subgroup Who Were Still in FAA ATC Work as of January 1973 (from Cobb, Young, and Rizzuti, 1976).

These attrition rates, which showed (1) no sex differences in the proportion of trainees who completed FAA Academy training, but (2) a percentage of women who subsequently left ATC work that was over twice that of male graduates, led to several studies of sex differences in job attitudes and attrition (Mathews, Cobb, and Collins 1975; Mathews, Collins, and Cobb, 1974a, 1974b). The category that accounted for most of the difference in post academy attrition rates between the sexes was family-related (marriage, children) attrition cited by about one-third of the women, a finding that agrees with results of other studies of different occupational groups (Mathews, Collins, and Cobb, 1974a).

Research on Military ATCS Trainees

CAMI has had a long-standing interest in the selection and training of military ATCSs reflecting the fact that former military controllers have always represented a prime source for the recruitment and selection of FAA ATCS trainees. As a result, the FAA has maintained awareness of the effectiveness of the controller selection and training programs in the U.S. Air Force, Army, Navy, and Marine Corps.

In 1965, representatives of the Glyno Naval Air Station, Georgia, visited CAMI for indoctrination regarding the screening and selection of applicants for FAA ATCS training. Naval officials subsequently discussed the possibility of using the operational battery of CSC tests for experimental administration and validation at Glynco. Since a number of policy reasons precluded this approach, Cobb suggested an alternate plan, subsequently accepted, involving the seven commercially published tests that had been validated in previous research with FAA trainees (Cobb, 1971). After completion of the Glynco study (Cobb, 1968b), officials of Keesler Air Force Base, Mississippi, asked that a parallel study be conducted on samples of Air Force and Army ATC trainees.

Although two of the seven commercially published tests which were administered experimentally to the military ATCS training schools failed consistently to correlate significantly with the training-performance measures of every group (CTMM Analogies and CTMM Coins), composite scores based on the entire seven-test battery correlated with the academic plus laboratory grades of every group at somewhat higher levels than measures based on any other combinations of the test scores. At the same time, in every instance, the validity of the "commercial seven-test composite" score was closely approximated by that of a composite measure reflecting performance on DAT Space Relations, DAT Numerical Ability, DAT Abstract Reasoning, and CTMM Inference (Cobb, 1971).

Both composite scores correlated significantly better with the training-course performance measures than did the Air Force's General Aptitude Index (GI) and the Marine Corps' Military Screening and Classification (MSC) test score, and at about the same level as the aptitude screening measures used by the Army and Navy.

On the basis of these results, Cobb concluded that the military services had the capability, if desired or required, of upgrading their

109.

screening of ATCS personnel without new or additional tests, by establishing higher minimum MSC requirements. Any changes in the military requirements would be of direct interest to the FAA since the agency has and will probably continue to select significant proportions of its ATCS trainees from among those applicants who have pre-FAA ATC experience and who also are able to qualify on the CSC ATC Aptitude Screening Test battery. As a byproduct of this research, Cobb was able to outline the relationship between military selection/training procedures and the screening and entry of ex-military personnel into the FAA ATCS program.

The Multiple Task Performance Battery

While paper and pencil tests have done an efficient job in the identification of individuals who possess the "elemental skills and knowledge necessary to become satisfactory controllers, they do not measure one kind of ability that is believed to be relevant for a good controller--the ability to perform several different tasks simultaneously (Chiles, Jennings, and West, 1972). A technique that could provide an objective, reliable index of the ability to time-share multiple tasks relevant to the ATCS job, was considered of value as an adjunct to existing selection devices and hence the CAMI Multiple Task Performance Battery (MTPB) was developed.

The MTPB was originally designed for the Air Force as a device for use in research on complex performance of the sort demanded of aircrew personnel (Chiles, Alluisi, and Adams, 1968). The elements of the MTPB were selected to provide objective measures of "psychological or behavioral functions" of relevance to Air Force operations. The functions measured include monitoring, information processing, mental arithmetic, visual discrimination, and interindividual interaction in the execution of procedures. These functions appeared to be relevant, not only to aircrew activities, but to complex jobs in general and to the job of the ATCS in particular. Moreover, the tasks, as used routinely over a number of years, have been structured to impose varying levels of demand on the subject with respect to the requirements for time-sharing. Good performers on the MTPB not only possess ability on the individual tasks but also are readily able to shift their focus of attention from one kind of activity to another without disruption of the ongoing process.

During the period from May 1970 to January 1971, exploration studies were carried out by Chiles, Jennings, and West (1972) to examine the potential usefulness of a performance measurement device as a predictor of the future performance of ATCSs. Five groups totaling 229 ATCS trainees were tested on the MTPB, and the predictor scores that were derived were correlated with ratings provided by FAA Academy instructors. Chiles and West (1974) checked the official FAA personnel roster as of January 1, 1973 to determine whether each trainee who participated in the earlier study was still actively employed with an ATCS job code. Although further validation is required to assess the potential of the complex performance battery approach to ATCS selection, direct support for the predictive power of the MTPB was obtained in the train study. Indirect support was also provided in that (1) the instructor ratings were predictive of the retention criterion and (2) the MTPB was a good predictor of the instructor ratings, (Chiles, Jennings, and West, 1972).

110.

SUMMARY

For over two decades the FAA Civil Aeromedical Institute (CAMI) has engaged in active research programs exploring most aspects of the selection of air traffic control specialists (ATCSs). The results of those efforts contributed directly to the establishment of ATCS selection standards employed by the Civil Service Commission (CSC). Early studies on the validity of aptitude tests for prediction of successful completion of Academy training led to a decision to use such tests for part of the CSC screening standard. Later studies led to the establishment of a maximum age standard of 30 years for entry into ATCS training. Most recent studies contributed to the validation of the current aptitude selection battery (made operational in 1981). CAMI researchers have evaluated the validity of existing standards, have examined numerous variables and alternative aptitude measures, and have provided a number of data-based recommendations that resulted in upgrading of the effectiveness of ATCS selection. This chapter has emphasized research on aptitude screening measures, attrition, age, prior experience, education, sex, minority status, military ATCS training, and the Uniform Guidelines on Employee Selection.

REFERENCE NOTES

1. Federal Aviation Administration, Office of Personnel, ATCS intake study, 1971.

2. Federal Aviation Administration, Office Personnel, Evaluation of the "150" program, 1973.

SUMMARY

For over two decades the FAA Civil Aeromedical Institute (CAMI) has engaged in active research programs exploring most aspects of the selection of air traffic control specialists (ATCSs). The results of those efforts contributed directly to the establishment of ATCS selection standards employed by the Civil Service Commission (CSC). Early studies on the validity of aptitude tests for prediction of successful completion of Academy training led to a decision to use such tests for part of the CSC screening standard. Later studies led to the establishment of a maximum age standard of 30 years for entry into ATCS training. Most recent studies contributed to the validation of the current aptitude selection battery (made operational in 1981). CAMI researchers have evaluated the validity of existing standards, have examined numerous variables and alternative aptitude measures, and have provided a number of data-based recommendations that resulted in upgrading of the effectiveness of ATCS selection. This chapter has emphasized research on aptitude screening measures, attrition, age, prior experience, education, sex, minority status, military ATCS training, and the Uniform Guidelines on Employee Selection.

REFERENCE NOTES

1. Federal Aviation Administration, Office of Personnel, ATCS intake study, 1971.

2. Federal Aviation Administration, Office Personnel, Evaluation of the "150" program, 1973.

111.

Chapter 5

RESEARCH CONTRIBUTIONS AT THE
OFFICE OF AVIATION MEDICINE (OAM)

Evan W. Pickrel

The Office of Aviation Medicine (OAM) is the principal Federal Aviation Administration (FAA) staff element responsible for biomedical and behavioral, and generally human oriented research, including research on selection, training, measurement of proficiency on the job, and performance assurance. Its mission is to apply knowledge gained from its own research program and from the work of others to the promotion of safety in civil aviation and to the health and safety of agency employees. In order to accomplish its mission, OAM conducts research on various aspects of the human element in aviation. Offices and services of the FAA, such as Air Traffic, Personnel and Training, Flight Operations, and the Technical Center (hardware research) identify their human-oriented problems for resolution through OAM research. As the research is accomplished, its products are provided to the requestors and assistance is given in the translation of these products into operational applications. Most of the research efforts are conducted at the Civil Aeromedical Institute (CAMI) located at the FAA Aeronautical Center, Oklahoma City, Oklahoma. They are supplemented by research contracts granted to other government agencies or to private institutions by CAMI or by the OAM.

The OAM has a small but uniquely qualified research cadre, with extensive experience in government and industry. This office annually sends out a formal request to the other offices and services to make known their aeromedical research requirements, and annually transmits these to CAMI in the form of a research program guidance and current policy statement. Aperiodic requirements also are received and resolved in accordance with the operational need and the research resources available. Members of OAM accomplish some of this research; unique examples include the highly publicized hijacker profiles for passenger boarding gate screening in defense against hijackers, and flight personnel tactical programs for inflight defense against those who manage to get on board (Dailey & Pickrel, 1975a,b). Many research requirements, including these unique examples, originate within the OAM itself.

Psychiatric screening. The Office of Aviation Medicine has an operational need to evaluate the health status of FAA personnel, and to determine the job relatedness of their medical needs and fitness for duty. To do this, it is necessary first, to determine the medical and psychological-psychiatric status of those personnel before they become employees, so that responsibility is not assumed for conditions that existed prior to job entry. Past behavioral histories are an excellent basis for predicting psychological-psychiatric fitness of Air Traffic Controller Specialist (ATCS) applicants, and the Office of Personnel Management (OPM) provides historical information, gathered by government agencies, including the

113.

Federal Bureau of Investigation, State Department, and Immigration and
Naturalization Services, for use when screening for this sensitive posi-
tion (ATCSs control military aircraft if the nation is under attack).
Military records also are obtained for this use when available. Samuel
Karson (1969), a member of the Washington staff, did extensive research
on personality tests, toward developing quantitative measures for use in
such screening, but under existing OPM and Equal Employment Opportunity
guidelines, government agencies are precluded from using personality
tests for selection purposes. However, the FAA has used the Cattell 16
Personality Factor Questionnaire (16PF) for a number of years as a diag-
nostic aid, as a case identifier for ATCS physiological-psychiatric
screening. When questionable cases are identified, these persons are
referred for extensive psychological and psychiatric examination, and
the results are used in a medical determination to reject or not reject
the applicant. It was estimated that profile analysis of the 16PF iden-
tified between one and two percent of all applicants who warranted closer
medical examination, and that half of these were found medically dis-
qualified for service (Colmen, 1977). Use of the profile analysis of
this measure has since been replaced by a customized key developed by
John T. Dailey for this (ATCS case identifier) application (See Chapter
17), but the statistics on number of cases identified and rate of medical
disqualification remain very similar to those experienced with the more
elaborate, detailed profile scoring procedures.

Development of new screening tests. Among the recommendations of a
1970 FAA management Air Traffic Controller Career Committee (Corson, 1970),
was one dealing with improvement of the selection process as a means of re-
ducing an unacceptably high attrition rate both in formal training and
during on-the-job developmental training. An extensive research program,
both within the OAM and by contract was initiated in response to that
need. Colmen (Milne and Colmen, 1972) identified available tests to in-
crease the predictive validity of the existing battery, and suggested
areas for construction of new tests. One of these involved filmed simu-
lation of air traffic moving across a radar scope, in response to which
the observer was required to predict any potential conflicts that appeared,
as soon as they were observed. This instrument, named the Controller De-
cision Evaluation technique (CODE), by Buckley and Beebe (1972), added
significantly to a composite for predicting on-the-job success, but the
equipment and space required for its administration generally were not
available in Civil Service Commission decentralized testing situations.
Therefore, test development procedures were initiated by OAM personnel
to derive measures of the same skills in a format that would meet Civil
Service needs. This eventually resulted in the development of a completely
new test, the Multiplex Controller Aptitude Test (See Chapter 15).

Prior experience in air traffic control work has been considered to
be a favorable factor in the selection of ATCS applicants, and those with
previous military controller experience have been a prime source for candi-
dates. Many other types of experience, such as private and commercial pilot,
work in communications, and even general work experience and academic train-
ing, have also been regarded as favorable factors. The Civil Service

114.

Commission used a Rating Guide to evaluate applicant claims of past experience, and the variety among applicant claims made them difficult to evaluate. Thus an operational need was recognized for the development of a quantitative measure to evaluate the usefulness of various types of experience claimed by applicants. The Offices of Aviation Medicine, Personnel and Training, and Air Traffic, organized a two-week workshop, held in Washington in June 1970, which was attended by 18 journeymen from all 3 ATC options, to address this problem. John T. Dailey and others instructed these subject matter experts on the purpose and specifications for an occupational knowledge test, and instructed them in the initial item writing. At that time over 300 items were prepared, and these were reviewed and edited by members of the Examination and Certification Section at the FAA Academy. OAM test construction experts used these items during development of initial and subsequent forms of the Occupational Knowledge Test (See Chapter 16).

Early identification of potential failures; reduction of post-Academy attrition. Between 1970 and 1976, ATC Academy pass-fail training was suspended while a new training program was being developed and evaluated. During this period, trainees who ultimately "washed out" of the program served between 18 and 24 months, on the average, before separation occurred. The rate of hiring of new applicants was between 1700 and 2000 persons annually, and attrition rates in the two to five year training program required to reach full performance (or journeyman) level ranged from 25% to 40%. Such high attrition rates were not only costly to the FAA, both in time and dollars, but also unfair to those who were failing. A major part of this attrition was attributed to lack of proficiency, the students' inability to demonstrate the skills and knowledge needed to progress satisfactorily and to succeed in training. An operational need was identified for early identification of potential failures, and it was expected that this could be done during selected phases of initial training, using appropriate methods to evaluate interest as well as potential. Thus it was decided to validate and re-introduce a pass-fail training program at the FAA Academy.

Personnel from the Office of Aviation Medicine, Personnel and Training, and Air Traffic reviewed the measures available for use in pass-fail evaluation and recommended modifications as necessary. The achievement tests that measured possession of academic information taught in the classroom were judged to be adequate, but a new test, the Controller Skills Test, was created by the OAM staff and Air Traffic subject matter experts to measure application of knowledge and skills during the first months of ATCS training (See Chapter 11). The major purpose of this test was to augment over-the-shoulder ratings of performance on laboratory criterion problems and instructor ratings of each student's potential for successful future performance as a controller. OAM personnel accomplished field validation of this new Controller Skills Test and other tests to be used for pass-fail purposes and also accomplished the criterion reliability and validity analyses necessary prior to introduction of the new pass-fail training program (Dailey and Moore, 1979).

The success of this new program is discussed by Boone in Chapter 9. Attrition rates for the two to five year training program from entrance to

the attainment of journeyman status have remained at approximately 25% to 40%; however, with selected phases of pass-fail training centralized at the FAA Academy, the average length of service before separation, for those who "wash out," has been reduced to five months. This up-to-one-and-one half year earlier identification of potential failures has provided an estimated cost avoidance of ten to twelve million dollars a year when hiring rates were 1700-2000 persons annually.

Development of a new ATC Selection Test Battery. Cost avoidance through early identification of potential failure was highly desirable; nevertheless, overall attrition rates of 25% to 40% have remained a serious concern. When the hiring rate was 1800 new employees (trainees) annually, a 24% attrition rate before reaching journeyman status equated to an investment loss of about $43 million each year. More detailed training and attrition costs have been presented by Rock, Dailey, Ozur, Boone, and Pickrel (See Chapter 21). Salary and training costs for the initial FAA Academy qualification training program have approximated $10 thousand for each student who entered, and with an annual enrollment of 2000 students, an average 30% attrition rate could be translated into a $6 million a year expense. The Civil Service Commission (CSC) test battery used for initial screening of ATCS applicants had remained unchanged since 1964. That test battery was composed of generalized, factor-pure measures; it was believed that tests customized to ATCS activities would provide a more precise basis for prediction of the criterion.

Applicant statements describing prior aviation-related experience were another element in the selection process, and there were weaknesses in the evaluation and weighting of experience as presented in the application forms. Improvements in those selection procedures seemed a feasible way to reduce FAA Academy attrition rates. Work cited above, initiated early in the 1970's, had resulted in development of the MCAT and OKT to aid in meeting these operational needs. As described in Part IV, these tests had been administered to all entering students at the FAA Academy (who had also taken the CSC tests) since January 1976, for test development and validation. The results of a series of prediction studies, which included these measures along with others, led to the development of a new selection battery that was adopted in August 1981.

Research results reported in these prediction studies (summarized in Chapter 18) supported the conclusion that a new test battery comprised of the MCAT plus two measures from the old battery, Abstract Reasoning (CSC-157) and Arithmetic Reasoning (CSC-24) provided consistently high correlation with the criterion measures. It was also shown that the OKT provided a better measure of aviation-related knowledge than the OPM Rating Guide then used to evaluate experience, and that it demonstrated a significant relation to success in the ATC occupation that added to the validity obtained from the test battery.

The final step was a study to validate a proposed new test battery composed of two parallel forms of MCAT plus CSC-157, CSC-24, and OKT. The sample used included 953 trainees attending the FAA Academy during the period of June 1978 through December 1978 (See Chapter 21). The MCAT

was the major component in the validities obtained with the new battery, and CSC-157 also contributed to the multiple correlation for predicting ATC performance criteria, but CSC-24 did not. Weighting the tests contributed to the validity of the battery. When annual training investment losses through application of the old operational battery were compared to those expected with application of this new test battery for selecting 1800 applicants, the potential cost avoidance was estimated to be in the order of $3 million a year. However, many thousands of applicants had already been screened with the old battery and were awaiting an opportunity to enter training. There would have been no early need for more applicants unless there had been an unusual event, such as a PATCO strike and dismissal of those who refused to return to work. This created an immediate need to open the register and screen more candidates and as a result the new battery was pressed into service shortly after the strike occurred, as discussed in Chapter 6.

117.

Chapter 6

ADJUSTMENTS IN THE AIR TRAFFIC SERVICE
FOLLOWING THE PATCO STRIKE

E. W. Pickrel

On August 3, 1981, the Professional Air Traffic Controllers Organi-
zation (PATCO) declared a strike against the Federal Aviation Administra-
tion (FAA), and there was a walkout of about 13,000 of the 17,275 con-
trollers actively engaged in control of air traffic. Some of the strikers
returned, but about 11,400 did not and were subsequently dismissed. This
left the air traffic control system with a significantly reduced capacity,
and there was a very immediate, real operational need to maintain the
system and rebuild the workforce. Past research, including ATCS selection
research, made a considerable contribution toward meeting that need. Some
of the activities that took place are described below.

IMMEDIATE WORK FORCE

The Air Traffic Service management and staff was composed of ex-Full
Performance Level (FPL) controllers who had elected to progress upward
within the management structure. Long before contract negotiation time,
they had been alerted to the PATCO intent to strike and the detailed pre-
parations that were being made. Management had prepared a hierarchy of
system contingency plans, with variations dependent upon the number of
persons expected to be available to operate the system. All management
personnel completed refresher training as necessary to renew their FPL
proficiency, so that they could be available to control air traffic.
With advent of the strike, these management people returned to the con-
trolling of aircraft, and six-day weeks became the rule. Vacations were
postponed. Sixty low-level activity terminals were closed and the non-
striking controllers of these terminals were transferred to busier facili-
ties. Approximately 800 controllers were borrowed fro. he Department of
Defense, to assist until the rebuilding of the workfor was accomplished.

POSITION RESTRUCTURING

One of the earliest adjustments was a restructuring of positions at
the larger facilities. This involved identification of three basic job
categories of work performed: flight data specialist, non-radar controller,
and radar controller. These newly structured positions are progressive on
the pay ladder. The flight data specialist position is important, but
does not involve traffic control directly and is not included in the
GS-2152 (Air Traffic Control Specialist) series. Employees who are hired
for this position, but fail to complete the necessary training are ter-
minated. Initially, unemployed pilots were hired to fill GS-7 and GS-9
levels, and borrowed military controllers also worked these jobs until
they were checked out in EnRoute sector or Terminal tower cab positions.
Approximately 1300-1400 of these have been established as permanent posi-
tions that formerly had been filled by developmental controllers. These

119.

specialists issue clearances to pilots, but do not formulate them; they
also handle flight strips and make sure that they are posted. They have
the opportunity to apply for FAA Academy training by taking the same examina-
tion that others take.

The maximum grade level for the nonradar controller ranges up to GS-12,
depending on the facility. Progression from nonradar to radar controller
is competitive, with initial promotion based upon satisfactory completion
of training prerequisites, as well as available positions and time-in-
grade requirements. Chances for advancement depend on the number of
vacancies available, job performance and qualifications compared to
other competitors, and in some cases willingness to transfer to other
locations. All three categories of work are staffed in Air Route Traffic
Control Centers and in those Terminal facilities that are equipped with
cab/radar. Other Terminal facilities require the use of only one or two
of the categories, because of facility size or specialized operational
function. This restructuring of jobs enabled the employment of persons
of lesser aptitude in positions below that of full performance level con-
troller. At the same time, it was efficient in that it freed FPLs from
lesser, non-control duties, and reduced the number of FPLs needed in the
system.

A system study of the National Airspace System (NAS) was initiated to
identify potential modifications to increase the efficiency of the total
system. An airways realignment and air route traffic control resectoring
was initiated to make departure-destination routes more direct, reduce the
distance and time that aircraft are on the network, and optimize sector
workloads. The pre-strike national NAS structure of 790 sectors was re-
duced to 655 sectors, with an expected further reduction to 550 sectors.
Some Terminal resectoring also is possible, and as each sector requires
staffing of radar scope and other positions, this has resulted in a con-
siderable reduction in manpower requirements.

FLOW CONTROL

Prior to the PATCO strike, all aircraft had unconstrained access to
the National Airspace System (NAS). As a result there were extremely
high controller workloads at dawn and dusk as aircraft were departing or
arriving at destinations; workloads during the remainder of most days were
low. Nevertheless, manpower levels were programmed to meet peak traffic
demands. If aircraft departures and arrivals could have been distributed
more evenly throughout each day, and departing aircraft held on the ground
to be released as space became available in the NAS, there is no doubt that
they could have been controlled all along by a considerably smaller work
force.

Computer programs and a central flow control facility were available
to assist in meeting this need. These had been created approximately five
years earlier, at the time of the OPEC fuel crisis, to help airlines avoid
costly fuel delays in holding patterns during flight. In the present crisis,
it was realized that control over all aircraft using the NAS would impose a

120.

dramatic increase in the demands placed on this facility, and it was moved from Jacksonville, Florida to FAA Headquarters in Washington, where greater manpower and communication capabilities were available. Its introduction resulted in a more even distribution of workload, with daily optimizing of traffic load in accord with available controller capability. Some air traffic had to be curtailed at the busiest facilities but, with deferred takeoffs to achieve a better distribution of the day's traffic, 83% of the pre-strike traffic could be accommodated by a workforce only 42% the size of the pre-strike force. Under the new system, the rate of operating errors was also 36% lower than that of the pre-strike workforce for a similar workload. This was a far more efficient system and it reflected the efforts of a highly motivated work force.

EXPANDED TRAINING

Only eight days after the strike was called, the first group of 144 students entered the FAA Academy. These represented appointments from the register of 9000 persons tested with the old Civil Service Commission (CSC) ATCS battery that had been used without revision since 1964. Plans were implemented for a new group of 500 students each month, a total of 6000 a year in training, until the end of 1983, when it was expected that the work force would again be up to full strength. The procedure followed was for FAA Regional Office personnel to select the names of applicants, in sequence off the top of the register, to fill vacancies in their facilities. Each selected applicant received notification of the available position with a description of the work and, if interested, was asked to proceed to the nearest facility of that type for interview and a medical examination. Following completion of these and a security check, the applicant was notified when an entry into training would be available at the Academy. To meet this new training requirement, the schedule of the FAA Academy was expanded to two shifts a day, and a contract with the University of Oklahoma was utilized as a source for hiring additional instructors. Most of those hired were retired FAA controllers and instructors who volunteered to help in the crisis.

A student's first few days (Phase I) at the Academy are spent in personnel processing. The Phase II activity, common to all options, is classroom training in fundamentals of air traffic control, such as weather, federal aviation regulations, and air traffic control communications. This is followed by non-radar training that is specialized for each option, and includes extensive laboratory training to prepare the student for immediate performance of duties at the facility. Under the new procedures, initial Academy training is completed at the conclusion of the non-radar phase. Radar training was removed from the Academy entrance training syllabus and the graduates go directly to facility non-radar control tasks.

After successful completion of this training, and a period of successful non-radar performance, these developmental controllers become eligible for radar training. They receive initial radar training on simulators at their home facility and are given a test to qualify for radar training at the Academy. If succesful, they return to Oklahoma City for the six week Radar Training Facility (RTF) course. If unsuccessful, they may remain in

121.

the non-radar controller status. Theoretically, these persons would
have been washed out in the old program. Now non-radar positions are
considered career positions for those who do not progress further. Since
many Academy graduates do not go to facilities that have radar, radar
training would be a waste of time for them and for the FAA, and this
restriction of RTF training for only those who move on to radar positions
has effected considerable savings of both time and dollars. Facility
training is customized to the location and position where each person
will be working, while Academy RTF training is generalized.

NEW ATCS SELECTION BATTERY

Although the number of applicants identified through the old selection
procedures and still on the old register was considerable, their availability
had been greatly reduced with the passage of time. There was a need for new
recruiting, and the Office of Personnel Management (OPM) began to plan for
autumn testing and screening of an unexpectedly large number of ATCS appli-
cants. This provided a timely opportunity to introduce a new selection
battery that had been under development since around 1970 (See Part IV).

OPM personnel were familiar with the new ATCS selection battery, inas-
much as some of them had participated in work on its validation. In order
to facilitate consideration, the FAA provided OPM with a detailed report
documenting the developmental research (Rock, Dailey, Ozur, Boone, and
Pickrel, 1982) and recommended the use of the new battery in fall testing
for the new register. In a review of the FAA report, the responsible OPM
personnel noted that "The report shows extensive criterion-related vali-
dity evidence for the battery. It also shows that the test battery did
have adverse impact for minorities and women in the research sample; how-
ever, analysis of fairness showed that the tests were not unfair to either
subgroup. In addition, the report describes an effective upward mobility
program which minimizes the effect of the adverse impact on selection
(Note 1). They recommended that the new battery be implemented as a
business necessity, and the Director, Office of Personnel Research &
Development, approved its use.

On August 17, 1981, a team of test development specialists in OPM's
Examination Services Division was assigned to prepare the new battery for
operational use. Test items from the final validated test forms were re-
viewed for technical adequacy, clarity, and statistical parameters. They
were then assembled into three parallel series or test batteries, based
on statistical and content parameters. This redundancy was necessary to
meet such requirements as retest of persons who request it, variation of
test content for control over compromise, and substitution of a different
battery should compromise occur. After the parallel forms were reviewed
and approved by an FAA representative, camera-ready copies were prepared
and 100 thousand copies of each series of the new tests were printed.
These were shipped directly to OPM examination points nationwide. Career
field brochures with test descriptions and sample items were prepared,
printed, and distributed to all applicants. Scoring procedures and key
clearances were prepared.

Following an announcement of the coming examination, 125,508 applications were received. During the first scheduled test period, October 15 to November 30, 1981, 47,345 individuals were tested. Scoring was carried out by the Staffing Services Center in Macon, Georgia, with key clearance on a sample of the first 500 answer sheets scored for each series administered. This procedure included preparation and review of item analysis data to identify questionable items, and review of each questioned item by subject matter experts, for errors in printing or content. Item and key approval or changes were made as appropriate to complete the key clearance. Scoring began approximately November 2 and was completed by the end of December.

The final rating for each applicant is based on the scores on the battery, consisting of two forms of the Multiplex Controller Aptitude Test (MCAT) (each form weighted 2) and the OPM Abstract Reasoning Test (weighted 1). In addition, up to 15 collateral points are given for performance on the Occupational Knowledge Test (OKT) and 5 or 10 additional points are awarded by law for veterans preference. The rating may not exceed 100 before the addition of veterans preference points. A newly created Special Examining Unit Staff (SEUS) at Oklahoma City was assigned the responsibility to add veterans preference points, to establish the ratings for the register, and to maintain and distribute registers for FAA personnel to use for the selection of new students in the GS-2152 occupational series. The old register was closed and the new register distributed on approximately the first of December 1981.

Comparative Performance, New versus Old Register Students

The first classes of students that entered the FAA Academy in 1982 included persons selected from the old register and the new register. Summary statistics for comparison of their pass-fail performance are presented in Table 1. The pass rate for 479 old register students was 43%, while the pass rate, for 965 new register students was 71%. This improved pass rate, which was expected based on the research reported in Part IV of this book, was very welcome, since 6000 students were scheduled to move through this training in the next two years. To illustrate using these figures, with a 43% pass rate, it would be necessary for 13,953 students to proceed through Academy Training before 6000 graduates were obtained. At a 71% pass rate, only 8,451 students would be required. The projected savings in time necessary to rebuild the workforce would be the primary return. The cost avoidance in Academy training, estimated at $10 thousand per student, would also be great. The comparatively small cost of creating these new tests, customized to better simulate situations calling for the skills needed by successful controllers, seems well rewarded by their successful performance.

NEW RESEARCH REQUIREMENTS

At the beginning of 1982, a workforce only 42% of the size of the pre-strike force was handling 83% of the pre-strike traffic, and there was considerable pressure to increase the traffic load even further. This smaller workforce was working long hours, six hours per day in active

Table 1.

FAA Academy Pass-Fail Performances, 1982
Terminal Groups 1-3, En Route Groups, 1-4
Comparison Between Old & New Register Results

Terminal	Total Number	Total Pass %	Males Number	Males Pass %	Females Number	Females Pass %	Minorities Number	Minorities Pass %	Non-Minorities Number	Non-Minorities Pass %
New Register	250	81	233	80	17	88	4	75	246	81
Old Register	99	68	78	55	21	57	7	57	92	68
En Route										
New Register	715	67	643	67	72	69	27	56	630	67
Old Register	380	37	301	38	79	30	14	50	134	42
Totals										
New Register	965	71	876	70	89	73	31	58	876	71
Old Register	479	43	379	45	100	36	21	52	226	53

124.

control of traffic (periodic breaks are needed when staring at spots of
light moving across a cathode ray tube), and 60-hour weeks were common
immediately following the strike. Much concern was expressed concerning
the stress and fatigue experienced by this work force and the possible
consequent effects on air safety. Management responded by cutting back
the work week so that it average 44 hours and cut back time on the boards
to an average of five hours per day. Independent studies by the Flight
Safety Foundation (1982) and National Transportation Safety Board (1981)
declared that the system was safe, but also indicated the desirability
of establishing a system to monitor worker stress and fatigue.

The Air Traffic Service had earlier requested the Office of Aviation
Medicine to develop a proficiency testing and quality assurance program
for Full Performance Level controllers, and the FAA Office of Personnel &
Training had requested a similar program for those in developmental train-
ing at the facilities. Joseph A. Tucker of the FAA Office of Aviation
Medicine developed models of such a system in concept papers, (Tucker,
Note 2). This included a tracking system and a data bank to describe
each individual's past and current capabilities, the use of a simulator
for periodic verification of job proficiency levels or need for remediation
by refresher training or medical assistance, and on-the-job monitoring to
assure quality peformance of the man-machine system and man in that system.
This latter quality assurance monitoring activity is designed to give each
controller more frequent, immediate feedback about his activities during
control of air traffic, to help him or her achieve, and to maintain quality
performance during his or her career. This is particularly important at
every point, as new controllers enter facilities and proceed through de-
velopmental training to certification at the Full Performance Level. The
expressed operational need for a system to monitor stress and fatigue is
met as these elements are identified with the total system effectiveness/
efficiency monitoring system.

SUCCESS OF THE ADJUSTMENTS

At the end of the first year following the PATCO job action, air
traffic seemed to be moving smoothly. The strike hurt the airlines, but
not as much as a recession that was already biting deep into air travel.
The airways were judged to be safer than before the strike, as traffic
was lighter and planes were spaced farther apart in the air. Concern was
expressed in Congress and by the aviation industry about continued restric-
tion on traffic, but the air traffic in the western third of the nation
again had been given unconstrained access to the NAS. The effect of the
PATCO strike was the accumulation of proof that facilities had been greatly
overstaffed, and plans were made to replace only about 7800 of the 11,400
people that were dismissed. As of August, 1982, 2300 new controllers had
completed their initial training at the FAA Academy and returned to their
facilities for on-the-job training. The Secretary of Transportation was
quoted to the effect that as more graduate, the last strike-related restric-
tions will be lifted by the end of 1983 (Washington Post, 1982).

A new NAS system plan was introduced in 1982, and an engineering de-
velopment program was initiated that will greatly increase the use of

125.

computers in air traffic control and eventually reduce the system's dependence upon man as an operator or active controller. Around the year 2000, computerized control, with backup computerized control in event of failure, is expected to be able to handle aircraft from their flight plan programmed point of departure to destination. The role of the man-in-the-loop (the controller) may only be to handle deviations from the flight plan, and to input changes to th computer in order to coordinate such changes, when requested by pilots. It has been projected that this new system will permit a work force half the size of that found at today's EnRoute facilities, to handle twice as much air traffic. The implications of such changes for research on testing and proficiency measurement are presented in Part V.

REFERENCE NOTES

1. Office of Personnel Management, Office of Personnel Research and Development. Preparation of operational Air Traffic Controller Examination Battery (Test No. 510, Test No. 157, and Test No. 512). Unpublished Report, 1982.

2. Tucker, Joseph A. Air traffic controller performance assessment. Federal Aviation Administration, Office of Aviation Medicine, Unpublished report, 1982.

PART II

JOB ANALYSIS AND CHARACTERISTICS
OF AIR TRAFFIC CONTROLLERS

This entire book is about air traffic controllers, but Part II presents information focused specifically on who they are, what they do, and about their careers.

John T. Dailey addresses the first two questions in Chapter 7, Characteristics of Air Traffic Controllers. This chapter discusses the job analysis of the various controller positions and then presents a fascinating profile of the men and women in the occupation. It also speculates about the probable impact of the projected automation of the air traffic system. In Chapter 8, Joseph A Tucker, Jr. addresses three important issues concerning the career performance of the controller. These are: aging, stress, and job performance assessment. The treatment of each issue complements the descriptive information presented by Dailey.

Chapters 19 and 20, in Part IV, also present normative data on the CSC and experimental tests, prior aviation-related experience, and eligibility for veterans preference credit for men and women and race-ethnic groups (the latter in Chapter 20).

S. B. Sells

127.

CHARACTERISTICS OF THE AIR TRAFFIC CONTROLLER

John T. Dailey

Much confusion regarding the Air Traffic Controllers and their problems has been caused by the fact that, technically speaking, there is no such thing as "The Air Traffic Controller." The air traffic control system is manned by a great variety of individuals, all of whom are coded by the manpower system as Personnel Code 2152, Air Traffic Control Specialists. These specialists are sub-coded according to their current skills and assignments and will have a succession of such designations during their careers in the Air Traffic Control System. Initially the ATC Specialist is designated as a Trainee, and may progress to the status of Journeyman or Full Performance Level. During a normal career a large proportion of the Code 2152 group become management or technical specialists assigned to ATC installations. It is misleading, in thinking of Air Traffic Control personnel, to consider only the individual at the console as a controller, and every one else in the system as "Management." Virtually without exception, the Air Traffic Control system, from top to bottom is run by personnel who have been trained and have proven themselves as full-performance-level controllers before going on to more responsible administrative and technical assignments in the system.

The Air Traffic Control Specialist may be assigned to a variety of installations, differing widely in size and function. The Specialist could be assigned to a Flight Service Station, a non-radar terminal, a radar terminal, a flight control center, or to a number of managerial, training, or liaison positions. The personnel in the system may move several times during their careers. A fairly typical progression might be from a small non-radar terminal to a small radar terminal and then on to a large radar terminal. Further progression might then be to a supervisory position, often involving a move to another installation, for further career advancement. The ATC center personnel are less mobile but they also often go on to fill supervisory and technical positions in the center or in supporting organizations.

The dynamics of mobility and career progression greatly complicate the study of "the Air Traffic Controller" over periods of time. The population of full-performance-level (or journeyman) controllers with different amounts of experience includes significantly different groups as more and more of the journeymen rise to more responsible or more technical assignments. Since it is quite possible that promotion tends, to some extent, to go to the more competent journeymen, any adverse job effects would be mixed up with the factor of progressive selection out of the status of full performance level. The problem is further complicated when the journeymen at a specific installation are studied, since they frequently transfer to higher graded journeyman positions at larger installations.

The situation is, in many ways, similar to the careers of men in the field of education. There the career flow is from rural schools to urban

schools, small schools to larger schools, and from the lower grades to
junior and senior high school. After a few years, a teacher may work
up to a classroom teacher in a well-paying large school. Further career
progression is then into administration, specialties such as counseling
and the like, or into college teaching. Only a small minority of men
starting out as public school classroom teachers make it to retirement
age as classroom teachers still in public schools, and those who do are
atypical. Consider how complicated it would be to study the effects of
twenty or thirty years of elementary school teaching on men teachers.
The complexities involved in studying the long-term effects of working
as a full-performance-level controller are even more formidable.

The situation at the ATC Terminal at O'Hare Airport at Chicago is
an example of some of these complexities and the serious misunderstandings
that can be caused by them. Several years ago many of the controllers at
O'Hare believe that a few years service controlling traffic at O'Hare
would "ruin" a controller and cause serious medical consequences. At
the time, there were almost no journeymen controllers with more than a
few years experience at O'Hare. However, when groups of journeymen from
O'Hare were studied in a cohort design, it was found that they had a very
high promotion rate to administrative positions at O'Hare as well as
other installations. It is likely that a record of controlling traffic
at O'Hare (because of its reputation as being number one in size) was a
strong competitive factor when bidding on a desirable promotion. It was
found also that the medical disqualification rate there was about average
for large terminals. Most of the "alumni" of O'Hare are "alive and well"
and filling key administrative positions and technical positions all over
the ATC system.

Currently many controllers are concerned that relatively few con-
trollers make it to retirement. However, in actual fact, very few who
were journeymen controllers twenty to thirty years ago did not advance
to administrative or technical positions by the time they retired. In
the future, however, a much higher proportion may serve an entire career
as a journeyman. Fortunately the ATC system provides progress to the top
pay grades for journeymen with no administrative or technical duties.

Studying the tasks of the air traffic controller is complicated by
the fact that the controller performs a very large number of apparently
critical tasks and it has been found to be extremely difficult to reduce
this number to a manageably small number of prime tasks. Various approaches
have been made to achieve a valid simulation of the tasks of the controller
and to quantify performance on the tasks. These have produced from approxi-
mately forty to eighty separate performance measures that show no strong
clustering that would enable reduction to a few prime measures. These
task analyses have been helpful in the establishment of training programs
for the controller, but have been of minimum value to the problem of deter-
mining how to select controller trainees.

JOB ANALYSIS

Through the years a number of studies have been made of the tasks and
skills of the air traffic controller. Among the first of these was a pair

129.

of studies in 1960 by Davis, Kerle, Silvestro, and Wallace. Their studies were oriented toward the training aspects of the job. The first study was based on interviews with training supervisors at twenty-four towers and twelve centers. From them were obtained detailed accounts of the current training programs, their problems, and ways in which the program could be improved. The second study was based on data collected through the use of highly structured interviews with over 350 controllers at more than 40 facilities. Empirical job descriptions were obtained of all the operating positions in the towers and centers. The study also resulted in curriculum suggestions regarding:

A. Standardization of training.

B. Areas of non-standard training.

C. Sequences of upgrade training and career progression.

D. Transferability of personnel.

One early comprehensive study of human factors aspects of air traffic control was made by Older and Cameron, in 1972. They presented an overview of human factors problems associated with the operation of the then current system and of future projected air traffic control systems. Descriptions were included of those activities and tasks performed by air traffic controllers at each occupational position within the current system. Judgmental data obtained from controllers concerning psychological dimensions related to these tasks and activities were also presented. The analysis included considerations of psychophysiological dimensions of human performance. This study described the role of the human controller in the then current air traffic control system, particularly as that role was expected to change as a result of the system's evolution toward a more automated configuration. Special attention was directed toward problems of staffing, training, and system operation. A series of specific research and development projects was recommended.

The report included a basic list of the skills of the air traffic controller, which were divided into three major categories:

I. Input skills, including visual display monitoring, visual monitoring (non-display), auditory monitoring, and reading.

II. Processing skills, including information organization, selecting among alternatives, and information storage.

III. Output skills, including recording, reporting, and control operations.

A massive series of studies of the tasks and skills of the air traffic controller was carried out by the System Development Corporation (SDC) between 1972 and 1975. There were eleven reports in this series. The first of these, in 1972, covered Phase II - Local controller, ground controller, and radar controller. An earlier, Phase I, study had demonstrated

130.

that it was feasible to develop objective performance standards and measures for air traffic controllers.

Task data, performance standards, performance measures, and rating scales were developed systematically and analyzed for the EnRoute Radar controller position. After a series of field evaluations, the final recommended scales, measures, and standards and the procedures for use of the system by first-line supervisors were presented for use in the air traffic control system. Detailed work on the Phase II controller flow diagrams was also reported.

Performance rating instruction booklets were developed for the Cab (Tower) controller and the Center controller. A package of procedures and materials was presented for "over-the-shoulder" ratings and performance standards. This package was based on performance standards for work elements, and included 53 standards for the local controller, 36 for the ground controller, and 38 for the center controller.

In 1974, a report was completed of the FAA Tower (Cab), presenting descriptions and flow diagrams of control functions. This provided detailed descriptions of the functions associated with four air traffic control positions in terminal towers (Local control, Ground control, Flight data, and Clearance delivery). Detailed flow diagrams were presented to give a clear presentation of the duties performed in each position on a day-to-day basis. The flow diagrams were based on information obtained through direct observation of controllers performing their duties, from existing manuals, and through interviews with off-duty controllers. The flow diagrams were reviewed at selected facilities throughout the country in order to assure that they would be nationally representative. In 1974 a similar report was completed for the Air Route Control Center, including a report on the controller over-the-shoulder training review instruction manual.

Another 1974 report included an Air Route Traffic Control Center controller extended performance rating instruction manual. This described 45 functions of the radar controller and 46 for the manual controller. Rating on a work sheet was on a seven point descriptive rating scale.

The series of SDC studies was completed in 1975, with four reports on the terminal option; Terminal Option Controller Performance Evaluation Report, Terminal Area (Tracon/Cab) Air Traffic Control Facility controller performance rating instruction manual, Terminal Area (Tracon/Cab) air traffic control facility controller training review instruction manual, and Terminal Radar Approach Control Facility (Tracon) descriptions and flow diagrams of control functions.

The series of studies by SCD was a truly massive analysis of the many skills involved in air traffic control and provided much valuable assistance in the training of air traffic controllers and in their utilization. However, it was of limited value in supporting work on the basic task of selection of controllers, because of the great number and complexity of the controller skills and functions reported.

131.

One of the best comprehensive analyses of the skills of the air traffic controller was that by Whitfield and Stammers, in 1978. They attempted to describe the major features of the controller's job and of the underlying skills, with particular emphasis on the processing of cognitive skills. Consideration was given to the influence of equipment development on the human skills required in relation to new developments in computer application. More sharing of tasks between men and computers was anticipated. These involved human factors problems in design of the controller-computer interface, particularly determination of the balance of responsibility between man and machine, or between controller and computer programmer. Optimal forms of computer assistance were considered, but it was stated that there was little real evidence for the optimum assignment of these decision-making functions. Consideration was given to the effects of automation on the attitudes and job-satisfaction of the controllers involved.

In 1978, Stammers briefly considered several aspects of human factors in air traffic control. The area of airfield air traffic control was introduced and potential increases in demands were outlined. The tasks involved were described as an initial approach to the problem of developing improved systems. The various methods of information collection and organization for such tasks were discussed. Possible future developments in systems were mentioned, together with their associated ergonomics aspects.

Hopkin, in 1979 briefly considered the problem of mental workload in air traffic control. After considering a number of approaches to workload measurement he concluded that none of them had proved to be helpful; all of the approaches tended to yield a large number of relatively independent variables, with little tendency toward meaningful clustering.

In 1974, Karson and O'Dell studied personality differences between air traffic controllers and the general population and also differences between male and female controllers. They found controllers to be somewhat superior to the general population on several scales of the Cattell Sixteen Personality Factor Questionnaire, but found few differences between male and female controllers. (Karson & O'Dell, 1974a,b,c)

Smith and Hutto, in 1975, studied the vocational interests of air traffic control personnel on the Strong Vocational Interest Blank. An Interest scale for air traffic controllers was developed, which differentiated between controllers and men in general, as well as men in other occupations. There were no substantial differences between the interest patterns for the three different ATC options. However, it was found that dissatisfied controllers scored lower on the overall air traffic controller scale than did satisfied controllers. In another similar study in 1979, Smith compared the job attitudes and interest patterns of air traffic and airway facility personnel, who are responsible for the installation and maintenance of the various electronic, electro-mechanical, computer, and engineering systems that provide communication, radar, and navigation services to the aviation community. Both groups completed the Strong Vocational Interest Blank and questionnaires concerning job satisfaction. There was general agreement about areas of job satisfaction and dissatisfaction. However, ATCS's reported more satisfaction than AFT's from

various aspects of the work itself and from salary, while AFT's were more satisfied with responsibility, working conditions, and Civil Service retirement. The AFT's were also more favorable to management than were the ATCS's.

From the selection point of view, the most important task simulation has been that developed by Buckley and Beebe (1972) at the FAA National Aviation Facilities Experimental Center. They produced a series of motion pictures, consisting of photographs of radar scopes with realistic, computer-generated air traffic. These were presented along with information such as course, altitude and speed of the aircraft in the form of a test, the Controller Decision Evaluation test (CODE). The purpose of the films was to evaluate comparatively various display configurations in terms of air traffic control decision making. It was noticed that the task (CODE) also produced a distribution of individual differences in socres among controllers which tended to correlate with their scores on other criteria. The attempt in the CODE simulation was not to reproduce or simulate actual air traffic control, but rather to abstract the essential decision-making processes involved, in a greatly simplified environment. The controller here does not control the traffic--rather a traffic situation is observed as it unfolds in a film as if the subject were watching "over the shoulder" of another controller. The task is to predict whether there will be violations of the separation standard and to write on the answer sheet the identities of the aircraft believed to be involved in conflictions. The test score is derived by comparing the predictions made with the conflictions known to occur in the film. Both the conflictions correctly detected and the "false positives" (i.e., pairs of aircraft predicted to conflict, which do not actually do so) are counted in the scoring.

A coded clock is shown continuously on the screen during the CODE test in order that the subject may record on the answer sheet the exact instant when the decision was made. Accordingly, it is possible to obtain a score on earliness of decision as well as accuracy of decision. An unlimited number of scores can be derived from the CODE test, but they are mostly variations or combinations of the following three basic scoreable factors:

1. Earliness of prediction of a confliction.

2. Number of conflictions detected.

3. Number of non-conflictions predicted as conflictions.

The first two types of scores differentiated strongly between experienced full-performance-level controllers and newly hired controllers with no prior ATC experience. However, the third type of score did not discriminate. The experienced controllers were able to detect a higher proportion of the conflictions and to detect them a longer period before they occurred. But the older controllers made more, rather than fewer, errors in predicting conflictions that did not really occur. This suggests that this sort of score did not really reflect a true core skill in air traffic control.

133.

Apparently, the trained, expereinced controller tended to call the "near confliction" cases as conflictions when he was making the decision well ahead of time. This seems to make sense under the conditions of the air traffic control system.

The central skill of the controller seems to be the ability to respond to a variety of quantitative inputs about several aircraft simultaneously and to form a continuously changing mental picture to be used as the basis for planning and controlling the courses of the aircraft. In the CODE film simulation of the air traffic control task, this was approximated by an integration of measures of the percentage of conflictions detected and the earliness with which they were detected. In this scoring each item (pair of targets) was converted to a nine-point stanine scale. Stanine values 2-9 were assigned on the basis of the earliness of correct prediction of the confliction. A stanine value of 1 was assigned if the confliction was not detected, was predicted after the 12-mile point, or was predicted but later crossed out. The stanine scale is a nine-point standard score with a mean of five and a standard deviation of two. Each step is equal to one-half of the standard deviation of the normative group. The normative group used was the total sample of journeymen plus new hires for towers, centers and Flight Service Stations.

The measures of speed of decision and accuracy of decision need to be integrated. One good way of doing this is to prepare a normative bivariate distribution where each point represents a combination of speed of decision and accuracy of decision. This distribution can then be sliced nine ways on the diagonal to produce a bi-variate stanine. This can be done for any pair of scores and is often a very useful way of combining two variables. This score was found to discriminate strongly between new controllers and experienced ones. The differences in means between the two categories was more than one standard deviation for a sample of 52 controllers.

In preparation for these decision-making functions, the air traffic controller must master a vast library of manuals, maps, regulations, letters of agreement, and similar publications that guide the planning and decision making. These cover such areas as weather, communications, navigation, geography, flight service, aeronautics, and flight regulations. All of these are constantly changing and the controller must keep up to date on them continuously. The controller must also be adept at reading maps, charts, and tables and performing mental arithmetic on the data read. The combination of these mental demands requires that the controller be a "perceptual-discrimination" athlete. This means that the job of the air traffic controller is beyond the capability of the average person and therefore a high degree of selection is required for the occupation.

CHARACTERISTICS OF AIR TRAFFIC CONTROLLERS

In order to obtain a comprehensive profile of the basic skills and aptitudes of two representative groups of air traffic control specialist trainees, a group of 52 entering trainees at the Air Traffic Control Academy was tested on November 19, 1968 with the various Dailey Vocational

Tests, and another group of 291 journeymen air traffic controllers was tested in 1970. These tests were designed to measure the potential of young people for a wide range of occupations. They measure a number of the most important skills found to be associated with preference for training, occupational choice, and performance. The sub-tests measure knowledge of electricity, electronics, mechanics, physical sciences, arithmetic reasoning, elementary algebra, vocabulary and spatial visualization. An electrical composite score was derived, based on scores from the electricity, electronics, mechanics, and physical sciences tests. A mechanical composite score was based on the mechanics and arithmetic reasoning tests. A scholastic composite score was based on the arithmetic reasoning, elementary algebra and vocabulary tests. These tests were modeled after the tests used by the United States Air Force and Navy for many years, for the selection and classification of technical trainees.

The test results were translated into percentile rank by grade 12 male norms. A percentile rank of 71 would mean that the average person in the group was better than 71 percent of the normative group, which happens to be a nationally representative sample of males in the twelfth grade. Conversely, a percentile rank of eleven would mean that the average person in that group would be better than only eleven percent of the normative group. By definition, the average score for twelfth grade males would be at the fiftieth percentile in this situation.

It was found that air traffic controller students vary tremendously in their relative average level of skill or aptitude in these various areas. They range from the eighty-sixth percentile in electronics and mechanics down to the fortieth percentile in elementary algebra.

Air traffic controller students were exceptionally high in mechanics and also extremely high in electricity and electronics. On the other hand, relative to this high level, they were at a far lower level in such things as physical sciences, arithmetic reasoning and elementary algebra. They were quite high in vocabulary and, of course, very high in overall score. They were also very high relatively in spatial visualization; indeed, they were the highest group of all in spatial visualization and also tied for highest with airframe repair students in mechanics. This is an extremely interesting finding. They may be extremely high in spatial visualization just because the air traffic controller aptitude tests stress measures of spatial visualization. However, the air traffic controller aptitude tests have no mechanical tests whatever, so this exceptionally high degree of mechanical aptitude must come by the indirect screening of the types of people who want to be air traffic controllers and the kinds of prior experience they have had. Nearly all air traffic controller students have had considerable prior experience with some aspects of the utilization of complex equipment in the communications, air control, or similar fields. This apparently is what generates their exceptionally high level of mechanical comprehension.

They are also extremely high in vocabulary, although the aptitude test did not include measures of vocabulary as such. Apparently man- f the

135.

indirect screening factors tend to generate a group that is very high in vocabulary. The group predominantly had little prior college training; many of the students did not complete high school, but acquired a high school diploma by examination. The low points on the profile in physical sciences, arithmetic reasoning, and algebra are typical for groups without college training.

In Figure 1 one can see the estimated average total score performances on the Technical and Scholastic tests (T & S) of a large number of groups. The scores on the T & S Test were estimated from the scores on the Information Test of Project Talent. In 1960 Project Talent tested 500,000 high school students and began a series of followup studies. In Figure 1 are shown students in various college or non-college groups, one year after graduation from high school. Major state universities were at about the seventy-second percentile, while all college students were at about the sixty-third percentile. Junior college students were at about the fifty-third percentile, whereas grade twelve students were at about the forty-second percentile.

Of primary interest are the non-college groups shown on the right of the figure. These range from waiter, at about the fourth percentile, to electronic technician, at about the seventieth percentile. As can be seen, the average air traffic control specialist student averaged at about the level of the electronic technician and was only slightly below the level of major state university students. They were above the average for all college students. These are students entering the academy and no attrition had occurred at the time of testing. One year later, undoubtedly their average score would have been considerably higher, since there would have been an appreciable amount of attrition in school and on the job, and past experience has shown that this attrition is highly related to aptitude level. It can be seen then that the air traffic controller group is an exceptionally high level group aptitude-wise, and indeed, is at a higher level than the average of all college students. They are considerably higher than the average senior college student.

Figure 2 shows the aptitude profile of an air traffic controller group, compared to norms for the male twelfth graders. Also shown on the chart are results for electronics students in specialty schools, airframe repair students in specialty schools, secretarial students in specialty schools and a group of broadcasting students in specialty schools. It can be seen that the air traffic controller group is a unique group. They have the very high technological aptitude of the electronics and airframe repair students, but also the high verbal level of the broadcasting students.

All of the groups shown in Figure 2 share a tendency to be relatively low in regular school achievement and this is a reflection of the fact that they are noncollege groups. The group of secretarial students was put in to show that such occupational groups have a profile that is exactly the inverse of that of the air traffic control group. This probably is one of the major reasons why women do not go into air traffic control in significant numbers. The reason is probably the same for why they do not go in substantial numbers into electronics maintenance, air frame maintenance or any other occupations or professions with a sizable technological component.

136.

Information Test Percentile	Estimated TST-Total Percentile	
99+	99+	Highest Individual
99	99	
		Three Highest Colleges
98	99	
		Highly Selective Colleges
95	97	
		Three Highest High Schools
89	89	
82	82	
		Major State Universities
73	72	All College Students
65	63	
		Junior College Students
56	53	
		Grade 12
47	42	Grade 11
		Total Noncollege Students
38	32	
		Grade 10
28	22	15-year-olds in High School
20	16	Grade 9
13	10	
		Three Lowest Colleges
9	5	
5	3	
3	1	
		15-years olds not in High School
2	1	
		Three Lowest High Schools
1	1	

Elec. Tech   Air Traffic Controller

Management business   Teller

Nurses aide   Trade management   Draftsman
Coast Guard(enlisted)   Marines(enlisted) Air Force(enlisted)
Navy (enlisted)   Airplane mechanic

Army (enlisted)   Radio/TV repair   Foreman
Electrician   Machinist   Plumber   Auto Mech.
General labor   Welder   Gas Station Attendant
Equipment operator   Carpenter   Assembly line
Butcher   Truck Driver (local)
Barber   Cook   Construction Worker
Bus Boy   Janitor   Dishwasher
Bricklayer
Porter

Waiter

Performance on the Project TALENT Information Test Total (R-190) of Groups
of High School Boys (Tested as Seniors in 1960) Who Were Members of Important Educational
or Occupational Groups in 1961 and Who Responded to a Follow-Up Questionnaire.
Selected groups of the total sample are also shown (Adapted from Dailey, 1964)
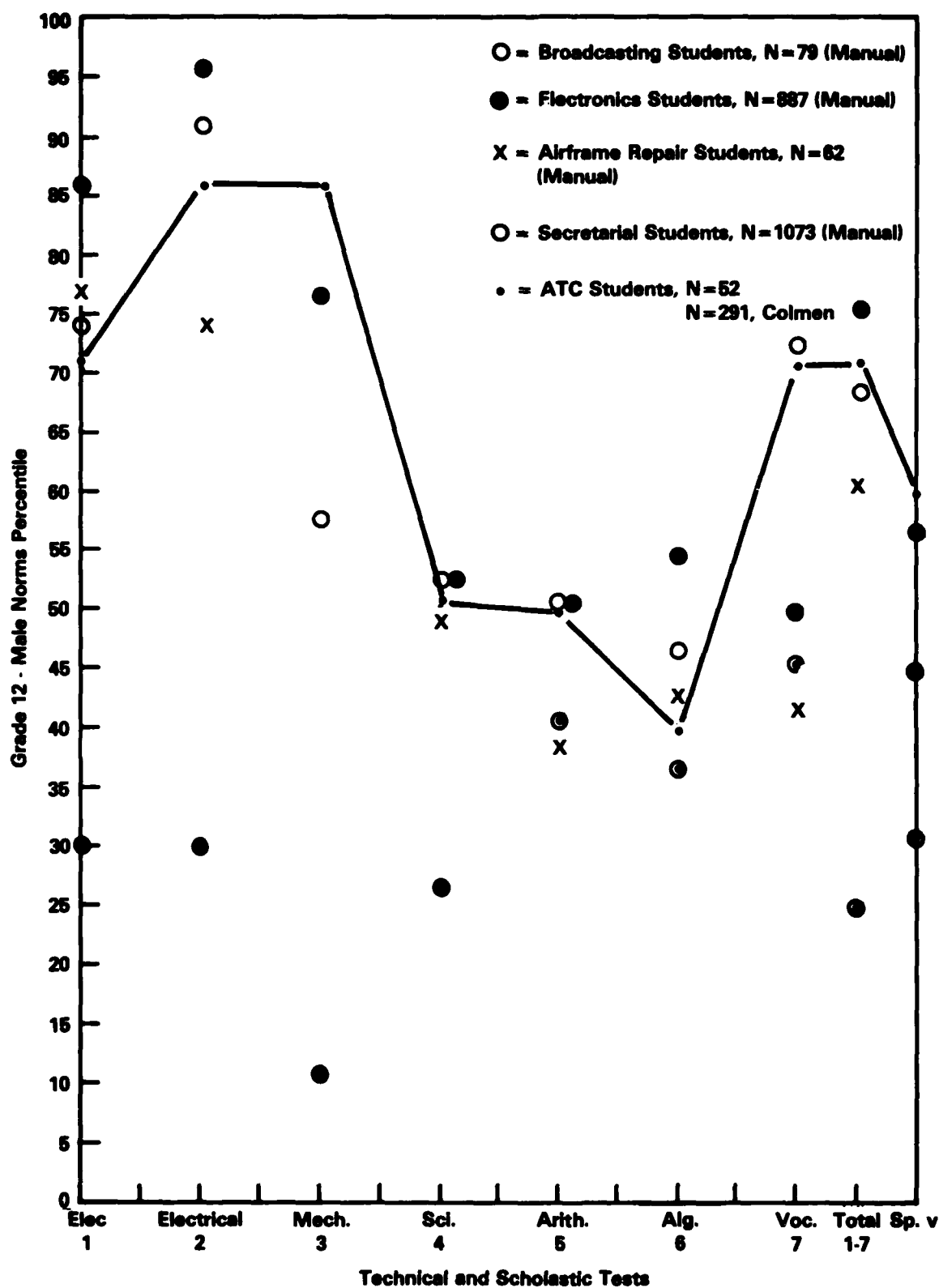
Figure 2. Aptitude Profile of Air Traffic Controller Group Compared to the Norms for Male Twelfth Graders and those for Various Specialty School Groups

138.

Figure 2 is of interest relative to the problem of cross-training air traffic controllers for other occupations or professions. It may be necessary in the future to convert some of the present air traffic controllers to other maintenance or computer programmer specialists or to other types of specialists than straight air traffic control. It appears from the profile of the air traffic controller students that there should be no difficulty in converting them into maintenance type jobs since they are very high in that area, showing a high degree of interest and basic preparation for any sort of maintenance or technological work. On the other hand, in converting them to any jobs requiring formal educational achievement, in particular basic arithmetic or simple mathematical skills, there may be trouble unless a combination of screening and preparation is utilized.

Experimental Testing of Journeymen Air Traffic Controllers

Test data are now available on a sample of 140 journeymen air traffic controllers, at the GS-12 and 13 levels. These data are shown in Table 1. This table also shows distributions of aptitude composite scores of those entering the Academy as trainees in 1961 through 1963 and the proportion at each aptitude level that survived to remain on the job six years later. It can be seen that, at any level, some of the trainees survive training and on-the-job-attrition and still remain several years later, presumably doing an adequate job. Those at the lower level of test aptitude who succeed in training and on the job do so because they have additional skills, motivation, and other characteristics not measured by the tests. At the very highest level of aptitude, all of the students were still on the job six years later. The proportion remaining after six years becomes smaller as the score levels looked at become lower.

Relative to the distribution of aptitudes in the general public, a composite qualifying score of 190 on the recently superseded OPM test (formerly the Civil Service Commission test) is extremely high and only a very small proportion of the general public could achieve that level. However, it can be seen that 80% of the air traffic controllers now on the job scored at a level of 190 or above and the median score for incumbents was 226. The last qualifying score of 210 (prior to adoption of the new selection battery), is very high. Despite the extremely high level of this standard, approximately two-thirds of the incumbents in the sample examined were above it. This indicates that air traffic controllers as a group, although they were not specifically screened on the Civil Service Aptitude Tests, are an extremely high level group on these aptitudes. They are at as high a level, aptitude-wise, as most professional groups that normally require college graduation for entry. This distribution of test scores by incumbent journeymen, as compared to the general population, tends to indicate that the tests were generally appropriate for use as a prescreening standard for air traffic control trainees and that a relatively high standard for this test seems to be reasonable. The new test battery adopted in October, 1981 represents an improvement over the previous battery.

In 1969 the author described the basic nature of the ATCS job as follows: "The air traffic control system is a man-machine system which

139.

Table 1

Cumulative Percentages of Air Traffic Controllers and the General Population at Various Levels of CSC Composite Score.

CSC ATC Composite Score	General Population	ATC Trainees 1961 - 63 (N = 893)	ATC Trainees Still On Job - 1968 (From '61-'63) (N = 501)	ATC Incumbents GS-12 & 13 (N = 140)
270+	.04	1.5	2.6	6.4
250 - 269	.40	6.3	9.4	26.4
230 - 249	2.3	18.7	24.6	48.6
210 - 229	9.2	36.6	44.4	63.6
190 - 209	24.8	54.9	65.6	80.0
170 - 189	49.7	74.5	84.6	89.3
150 - 169	74.6	85.6	93.0	96.5
Below 150	100.0	100.0	100.0	100.0

140.

places primary reliance on the human component, and the human operator is often the limiting factor in system output. The air traffic control system violates the first tenet of man-machine design—that is—'to design the system so the average man with an average amount of effort can carry out his part in the system.' Unfortunately, there seems to be no feasible way to do this and so the system requires a controller who is highly selected and highly trained. The controller bears a heavy load on his memory and his ability to keep many things in mind while arriving rapidly at well-reasoned solutions to complex problems under conditions of stress. Controlling air traffic is not necessarily a stressful activity, but in situations with a high traffic density, the job of the air traffic control specialist becomes stressful because it requires pushing man's capability to its limit to maintain continous peak performance.

This description is still valid as of 1982, but according to current research and development plans, the job and tasks of air traffic control may be vastly different within a few years. The plans call for automation of various controller tasks, one enhancement at a time until the job becomes mainly that of a passive observer whose main function is to stand by, ready to take over in the rare event of a system failure.

It appears that the anticipated stepwise automation would cause a gradual withering away of the air traffic controller specialty, to become a completely different and far less demanding occupation.

If this occurs, then the main problems of the occupation would become those of maintaining alertness while doing very little active work, instead of the present problems of coping with the stress of a heavy and demanding active workload.

It is believed that such a semi-automated air traffic control system as that planned for the next few years should continue to give the air traffic controller an active role in planning and managing the traffic for which he or she is responsible. The planned automation should be supporting this role of active controller. This may require some minor modification of the planned system in terms of how much active work the controller is expected to do. It would be desirable that this participation be as much as is reasonable rather than as little as possible.

It is doubtful that the planned automation during the next few years will make it possible to control more traffic with fewer controllers, where traffic is dense and more complex. However, it might well make it possible to permit more traffic in a given restricted airspace, if properly implemented.

141.

Chapter 8

AIR TRAFFIC CONTROLLER PERFORMANCE:

THREE CURRENT ISSUES

Joseph A. Tucker, Jr.

The Federal Aviation Administration (FAA) is the primary employer
of Air Traffic Controllers and has management responsibility over all
aspects of that employment including selection, orientation, training,
retention, progression and retirement. An important aspect of this
responsibility involves the assurance and periodic evaluation of the
continuing readiness to perform of each air traffic controller. That
responsibility is met operationally through selected missions of the
FAA such as these assigned to the Office of Aviation Medicine (OAM)
and Civil Aeromedical Institute (CAMI). This chapter reports on three
concerns that relate to job performance--aging, stress, and job per-
formance assessment. The data referred to were taken from studies
conducted prior to the Professional Air Traffic Controllers (PATCO)
strike in August 1981. That action does not affect the findings of
this chapter, but may affect future projections to some extent since
it has prompted changes in operating procedures and management practices.

Aging

The facts about air traffic controller aging and retirement present
a varied picture. The information presented here is from an Air Traffic
Controller Loss Study prepared in August 1980, by Headquarters, FAA
(Van Vuren, Note 4). A well documented fact is that controllers' careers
are longer than other federal careers. The length of a controller's
career is related to the retirement option selected. Over the decade
of the 70's, the average controller selecting the normal retirement
option worked to age 60, about two years longer than the average of
other FAA employees and two years less than the average U.S. Government
employee. Disability retirees averaged about 23 years of service and
retired at an average age of about 47 years. Controllers who selected
the Early Out option averaged about 28 years of service and retired at an
average age of 52. Up to 1975, most retirees selected the normal retire-
ment option. Between 1975 and 1981 the disability option was selected
most frequently. The current trend is such that it is projected that by
1985, the normal retirement choice will be most often selected, followed
by ATC Early Out and Disability, in that order. But the fact remains
that the Air Traffic Controller career within FAA is one in which con-
trollers evidence strong staying power, both as to average retirement
age and length of service. These data have been stable for ten years.
In fact, average length of service has increased one full year since
1971. Air traffic controllers enter service at an age younger than most
federal employees and a very large percentage make it a working life career.

Age of entry into service has been shown to have a strong inverse re-
lationship to successful performance in training and subsequently on the

143.

job. Research by Cobb, Mathews, and Nelson (1972) documented that the attrition among new selectees who were in their thirties or older was so high that FAA management limited selection for the EnRoute and Terminal options to applicants under age 31. This implies that air traffic control is a young person's occupation. A report by Cobb (1968) in the mid-and late 1960's supported the entry age cut-off by demonstrating that the job proficiency of full performance-level ATCS's or journeyman-level controllers generally tended to decline progressively after age 40. Evidence based on job performance ratings, that this decline has a significant detrimental effect on system performance, is not available. There is medical evidence of physiological changes with age, as is to be expected. Yet Booze (1978) concluded, after studying the morbidity experience of over 28,000 air traffic controllers over a ten year period, "that experience does not appear excessive when compared with the experience of other outside groups studied, except for psychoneurotic disorders. Additionally, a lack of association between disease occurrence and occupation is observed in the data correlating disease occurrence with length of service and age."

Pickrel (1979), in validity studies undertaken to support the use of a pass-fail criterion at the FAA Academy, documented the fact that air traffic controller performance improved for the first several years on the job. Decrements associated with continuing service occurred primarily among controllers assigned to administrative duties. Aging, per se, was not a factor.

A study by Thackray and Touchstone (1980a,b), on the effect of age on the ability to sustain attention during performance on a simulated radar task, showed that mean target detection time, errors of omission, and errors of commission increased significantly with age; performance impairment occurred earlier in the 2-hour experimental session, with increasing age. Physiological measures of visual scanning activity and skin conductance level, as well as subjective measures of fatigue, bordeom, and attentiveness all failed to provide adequate explanations for the greater decline in performance with age.

J.W.K. International (1981) conducted a study of Flight Service Station Specialists (FSSS) to provide information needed for a determination of whether Early Out policies for EnRoute and Terminal air traffic controllers should be extended to the FSS option. Emphasis was directed toward the effect of age and length of service on job performance. The study ended with the conclusion that: "Where there is a negative relationship between age and job performance as measured in this study, it explains only a small portion of the variance in FSS performance. It is fair to infer that chance or other factors not measured in this study could have an equal or greater effect on FSS job performance." (p. 116)

Thackray's work, some medical findings, and evidence of performance decrements suggest that research into the effect of aging must continue, but it must get closer to the actual career experience of the air traffic controller. There is much diversity of activity among the three air traffic control options. There are seven Flight Service Station positions

and three levels of facilities. The Terminal positions include radar
control, VFR, and movement of traffic on the ground. The Air Traffic
Control Centers similarly have a variety of positions. In addition,
workloads vary from facility to facility and from shift to shift.
It is important to determine whether the aging controllers are being
used selectively in a manner significantly different from the utiliza-
tion of younger controllers. In other words, if there is an implicit
recognition of performance decrement with increasing age among air
traffic controllers, is it being managed within the system in some
manner?

In addition to the on-going research at CAMI and other FAA agencies
on the aging process, several other approaches to research may be in
order. The Personnel Management Information System should be evaluated
as to how well it is tracking key career decisions by controllers and
should be improved where necessary. Longitudinal studies are likely
to be more informative than the cross-sectional studies so often used
in this area. And, finally, the writer recommends strongly that con-
sideration be given to use of the modern, sophisticated Naturalistic
Inquiry Methodology (Guba, 1981), which would be aimed at finding out
what really does happen career-wise among controllers as they build up
lengthy periods of service. The lack of conclusiveness of the research
on aging at this point suggests that we might not yet have identified
the right questions to ask. The Naturalistic Inquiry Methodology could
lead at least to the identification of the "right questions."

Stress

Since the experience of stress is subjective and personal, it is
very difficult to state a concise, objective operational definition of
the concept. The experience "stress," has physiological, emotional
and intellectual correlates, all of which help to define the scope of
the concept. The word "stress" has many meanings in the English language.
The definition in terms of "emotional or intellectual strain or tension"
(meaning six in Funk and Wagnalls, 1963), is perhaps most common. Within
the context of that definition, strain is "overexertion," and tension is
mental strain or intense nervous anxiety.

The job of air traffic controller is one that requires cognitive
facility and emotional stability. It is an active, participative job
rather than a passive one. However, it is not a physically demanding
job. The work itself does not lead to physical overexertion and physical
fatigue. Consequently, when terms that have physiological overtones such
as fatigue and stress are used for the air traffic controller job, it is
assumed that they have an emotional origin. It is less often recognized
that they also have an intellectual cognitive origin.

The workload of an air traffic controller can vary during a shift,
or as a function of location, from low activity to high activity. Emo-
tional tension can be associated with all levels of job activity and the
question of what level of activity in air traffic controlling correlates

145.

with the minimum of emotional tension is an important human factors question.

Stress of an intellectual, cognitive origin can arise from an informational processing overload. In the following discussion, stress is considered first, in terms of information processing overload and then, as anxiety.

Information processing. An important article by Finkelman and Kirschner (1980), defined stress in information processing terms. They assumed that the air traffic control tasks place unusually high information processing demands upon controllers for extended periods of time, so that they must work close to the limits of their channel capacity. The summary of their article includes the following interesting statement.

"The effort required to process information, maintain continuous concentration, and render timely and reasonable decisions is likely to be very stressful. Although stress-related performance decrements would not be acceptable in the typical air traffic control situation, the effects of stress may manifest themselves in social and family relationships and in physical and mental health. It is possible that laboratory measures of information processing (such as the delayed digit recall subsidiary task) could be used to evaluate reserve capacity and thereby predict the ability to cope with stress. Air traffic controllers with higher channel capacities may be less likely to make errors under conditions of stress and less likely to suffer the physiological consequences associated with high information processing demands." (p. 161)

The current lack of empirical support for common sense conclusions that one can draw about the role of information overload upon air traffic controller stress and illness should not deter human factors research based on informationprocessing theory. Channel capacity, an obvious analogy, is the key concept. The trained air traffic controller is assumed to have an ability to process air traffic that is relatively stable in each controller, but varies as to capacity from controller to controller.

Each is assumed to have a reserve capacity that can supplement the normal capacity, and this capacity is also assumed to vary from controller to controller. Under conditions of heavy workload, the controller with a relatively low normal capacity would have to call upon this reserve capacity sooner than a controller with a high normal capacity. The low capacity performer would be likely to experience stress sooner and experience it longer, and possibly be more potentially prone (predisposed) to a performance error than the high capacity performer.

The theory is helpful in explaining the relationship between stress and job performance. Emotional stress arising from sources other than the work itself, such as family, social or non-task-specific work environment situations, is assumed to degrade channel capacity, especially the "reserve capacity." This implies that the controller's capacity to perform

146.

satisfactorily may be reduced by stress which has an effect (for him) of raising the work load. In this instance, emotional stress is assumed to function as an independent variable. A linear relationship is implied but may not be warranted, as is pointed out later.

The theory also implies that a continued demand for production under conditions that restrict channel capacity can lead to stress or to exacerbation of existing stress. In this case, stress acts like a dependent variable, but is a moderator of the true dependent variable, performance (or output). If it were assumed, as is often done, that stress, as a dependent variable, can interact directly with job performance, then three possibilities exist.

One, a high level of emotional stress can lead to potentially poorer job performance. The evidence for this is controversial at best, even though the hypothesis seems to make common sense.

Two, a moderate level of emotional stress causes increased mental alertness and facilitates job performance. There is evidence that this may be the case, although words like "stimulating, energizing," etc., would be preferred, rather than stress.

Third, there may be little relationship between emotional stress and job performance when each is viewed as a dependent variable. Stress may affect family and social matters and possibly long term commitment to a career, but not necessarily the work itself.

A preferred explanation of the effect of stress arising from high workload is that though it arises from the workload and is, then, dependent on it, it feeds back upon channel capacity in the same manner as other, independent (e.g., emotional) sources of stress and can degrade channel capacity. From that view it acts as an independent variable upon job performance, even though it may remain an intermediate dependent variable for other effects upon the controller (Danoher, 1980) until some plausible "theory of action" can explain its effects.

Previously, I referred to intellectual or cognitive stress in relation to information processing. This leads to the hypothesis that a heavy workload leading to excess demand on channel capacity can cause cognitive stress that degrades performance, without emotional stress as a causal counterpart. Overload, in this case, creates a situation in which the individual may perceive his response capacity to be inadequate, and may induce a state of stress. Sells (1970) emphasized several important points related to this process: (1) Factors other than overload (e.g., distraction by personal problems, illness, or external factors in a situation), as well as the demand of the task, or a combination of these, may cause an individual to perceive himself unable to perform as required; (2) The consequences of failure to respond effectively depend on the personal involvement of the individual in the performance, which is determined mainly by the importance of successful performance to the individual. Thus stress intensity depends on the importance of individual involvement and the individual's assessment

147.

of the consequences of his inability to respond effectively to the situation. Sells stated as advantages of this definitional paradigm, first, that it provides a general concept of stress, independent of specific mechanisms, thus embracing a wide range of situations, personalities, and events; second, that it indicates an unrestricted range of measures to mitigate stress, by conditioning, training, equipment and system design, preparatory communication, warning, and perhaps even manipulation of costs and rewards; and finally, that it requires no special concepts of stress behavior, but rather utilizes established principles of behavior while providing a new principle to distinguish stress from other phenomena of human behavior, such as emotional arousal, fatigue, fear-anxiety, and others.

These conceptions have implications for selection, training and performance assessment and are researchable. Laboratory studies by Thackray (1981) at CAMI support the existence of information overload phenomena. Research into selection instruments that may detect tolerance for high information overload seems justified, particularly, at a time when increases in air traffic controller workload are forecast.

In summary, stress, from an information processing viewpoint, can both affect an individual's job performance and be affected by job demands. A plausible conceptual relationship to "channel capacity" can be postulated and tested. The relationship between stress and other variables may be non-linear and require careful plotting.

Anxiety. Anxiety has been defined psychiatrically as "a tense emotional state characterized by fear and apprehension regarding the future." (Funk & Wagnalls, 1963) As a psychological construct it has been thoroughly developed, tested, and correlated to many trades and professions, including air traffic control, and is considered to be a form of emotional stress. Frequently used assessment instruments include the Manifest Anxiety Scale (Taylor, 1953) and the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1969), both of which have rationales based on the Hull-Spence learning theory (Gaudry and Spielberger, 1971).

Prompted in part by recommendations in a NTSB (National Transportation Safety Board) 1981 special investigation report, the Federal Aviation Administration is proceeding with the development of a fatigue/stress detection program for air traffic controllers. Although the NTSB found in its October 1981 report no evidence of performance decrements for the then current work force, that agency recommended that FAA monitor stress and fatigue because of increased work hours. The NTSB also identified an immediate need for improved journeyman air traffic controller performance assessment procedures. It is likely that ultimately a stress/fatigue detection program will become a component of an overall performance assessment system.

Tucker (Notes 2 and 3), in two internal Office of Aviation Medicine Memoranda, proposed broad requirements for both a stress/fatigue detection program and a performance assessment system. David J. Schroeder (Note 1) at CAMI, is doing in-depth research into the stress/fatigue phenomena.

148.

A problem that continuing research and development must address is the ambiguity of the concept of stress as it is used in relation to air traffic controller performance. Spahn (1977), in a study by MITRE Corporation of the human element in air traffic control, reported that stress as categorized in the System Effectiveness Information System (SEIS), has 7 sub-categories that are disparate in meaning, varying from illness to job attitude. Stress never was identified as a direct cause of a system error in over 1200 system error reports, and was identified as a contributing cause in only five; age and length of experience were not factors contributing to systems errors (Kinney, Spahn, & Amoto, 1977). The discussion of stress continues under anxiety, below.

Smith (1980) has reported on a decade of research concerning stress, anxiety, and the Air Traffic Control Speciality. Smith's ten studies included attitude surveys, the State-Trait Anxiety Inventory, other anxiety inventories and physiological measures. Smith found air traffic controllers, in general, to score low on trait anxiety. There is little reason to assume that trait anxiety, as a personal characteristic of air traffic controllers, acts to degrade channel capacity, with a possible effect on job performance.

State anxiety, on the other hand, is sensitive to shift length and work loads. State anxiety scores are higher toward the end of the shift, a condition true of many professions. These scores also tend to be higher on night shifts as opposed to day shifts. However, the state anxiety score levels observed among air traffic controllers were too low to suggest impairment either of job performance or of channel capacity. Smith concluded that:

"There is little evidence to support the notion that ATCS's are engaged in an unusually stressful occupation. That is not to say that ATCS's never encountered unusual stress on the job; however, it does appear that this is the exception rather than the rule. ACTS's appear both well qualified and well suited for air traffic work. The demands of air traffic work do not appear to place unusual stress on ATCS's; this professional group appears quite capable of handling requirements of the job without distress. The notion that this occupational group is being pressed to the psychological and physiological limit is clearly unjustified."

A study conducted by the Institute of Social Research, University of Michigan (1975) compared stress factors in 23 different occupations. Although largely subjective and based on only about 100 men per occupation, this report appeared to support the position that air traffic control is not necessarily the most or even a uniquely stressful occupation. Nevertheless, the study concluded that in regard to the demand for mental concentration on the job, train dispatchers and family physicians were rated with ATCS's at the highest levels.

Much depends on the criteria chosen in assessment of stress. In fact, the report asserted that "if one were to peek at the most stressed occupational groups, they would tend to be the machine-paced assembly line

workers," an effect of boredom, dissatisfaction with the work load, and dissatisfaction with the job as a whole.

This brings up the topics of boredom and monotony in relation to air traffic control work, for there is anxiety among human factors specialists that increased automation may produce such an effect for air traffic controllers (Nealey, Thornton, Maynard, & Lindell, 1975). Thackray (1980) studied this matter recently at CAMI, and concluded that:

"The available data offer no support for the belief that boredom, monotony, or under-stimulation per se produce the syndrome of stress. However, monotony coupled with a need to maintain high levels of alertness, which might exist if controllers lacked sufficient confidence in an automated system, could represent a combination capable of eliciting considerable stress."

Performance Assessment

Job performance is the principal basis for decisions concerning controller retention and progression. Job knowledge is critical but does not appear to be a satisfactory criterion for distinguishing between good and poor performance.

Controllers evaluate each other. In training, the evaluations are based on laboratory exercises, instructor ratings, and skills tests. When proficiency is reached, supervisors' ratings are the primary indicator of acceptable performance.

Elimination rates at the FAA Academy have varied from 20% to 40%. Subsequently, in the field, the elimination rates may reach 20%. However, elimination percentages aside, the fact is that controllers believe that they can distinguish between good and poor performers and do so using the job or job-related data in making pass-fail decisions. The impressive predictive validities reported in Part IV of this book testify in part to the reliability and validity of the proficiency assessment measures.

For over ten years, research and development concerning performance assessment of journeyman air traffic controllers has been directed toward the development of more objective measures and the reduction of the dependence on subjective, "over-the-shoulder" ratings by supervisors. This work has also contributed to the establishment of aggregate criteria for research on the validation of selection and progression instruments. Mies, Colmen, and Domenech (1977) developed an aggregate criterion consisting of measures of training performance, job performance, and attrition. Buckley, House, and Rood reported in 1978 on nearly a decade of research that led to an approach to objective measurement of radar control performance of air traffic controllers by means of air traffic control simulation exercises. The Institute for Defense Analysis, in a report on the training of air traffic controllers (Henry, Kamrass, Orlansky, Rowan, String, & Reichenback, 1975), made a strong recommendation that the FAA develop objective job performance assessment procedures. This report pointed to the work of Buckley and associates as evidence of its feasibility.

150.

Kinney, Spahn, and Amoto (1977) reported on some innovative studies
of system errors and pointed out the difficulties of assessing the human
factor component of the air traffic control system in the absence of ob-
jective normative data. Their report contained recommendations for assess-
ment and control that have not been presented by other investigators. Boone,
VanBuskirk, and Stein (1980) reported on the continuing research at the FAA
Academy to develop objective performance measures for use at the new radar
training facility. Mohler (1980), in a report dealing with the health and
safety of air traffic controllers, recommended that the FAA "develop a
more meaningful recognition and awards program for controllers that is
directly related to quality performance." Pickrel (1979) has used per-
formance measures to establish pass-fail cut-offs for Academy training
and has urged the use of Controller Skills Tests to support the reliability
and validity of laboratory scores. Collins, Boone, and VanDeventer (1980,
See Chapter 3) have discussed the objective criterion problem in their
report on contributions by the Civil Aeromedical Institute. Tucker (1981)
has been investigating the feasibility of using paper-and-pencil micro-
simulations for objective skills testing both at the FAA Academy and at
air traffic control facilities; a further discussion of this work is pre-
sented in Chapter 12.

As an aftermath of the PATCO strike, the FAA is now proceeding to
investigate the feasibility of developing an objective ATCS performance
assessment system. The NTSB report concerning the safety and efficiency
of the air traffic control system published in December 1981 presented
40 conclusions that include the following: "(18). The foundation of
adequate ATC system capacity is the capacity and proficiency of the indi-
vidual controller. The over-the-shoulder training evaluation is not pro-
ductive and should be replaced by a more meaningful evaluation program."
These conclusions point to the need for an objective assessment system.

Currently, the Office of Aviation Medicine has been assigned the task
by the Director, Air Traffic Service to undertake a formal program "as soon
as possible because of the urgency associated with the rebuilding of the
air traffic control system" (Van-Vuren, Note 5). Tucker (Note 3) has pre-
pared a preliminary statement of the scope of such a formal program. This
work will proceed immediately, based on a sound research foundation. How-
ever, the determination of what constitutes mastery among air traffic con-
trollers, and the differentiation of the best controllers from the acceptable
ones awaits a theory of cognitive behavior that is descriptive of how the
cognitive repertoire of the controller is used in the control of air traffic
(Hopkin, 1980).

REFERENCE NOTES

1. Schroeder, D. *Sources of Stress*. Regional Flight Surgeons Conference. FAA Office of Aviation Medicine, Seattle, Washington, December 1981.

2. Tucker, J. A. *A fatigue/stress detection program*. Inter-office Memorandum, FAA Office of Aviation Medicine, Washington, D. C., December 1981.

3. Tucker, J. A. *Air traffic controller performance assessment*. Inter-office Memorandum, FAA Office of Aviation Medicine, Washington, D. C., January 1982.

4. Van Vuren, R. J. *Air Traffic Controller Loss Study*. Unpublished Report, Federal Aviation Administration, Washington, D. C., August 1980.

5. Van Vuren, R. J. *Controller performance skill testing*. Inter-office Memorandum, FAA Air Traffic Service, Washington, D. C., December 1981.

PART III

MEASUREMENT OF AIR TRAFFIC CONTROLLER PERFORMANCE

Accurate measurement of controller performance is essential both to
the validation of selection tests and to operation and management of the
air traffic control system. Since the focus of this book is on controller
selection, controller performance measures are generally discussed as
criteria, to reflect their role in validation research. Nevertheless,
other uses are mentioned throughout the following chapters.

The availability of reliable and valid criterion measures is criti-
cal to the empirical determination of the predictive validity of any
selection instrument. Validity is indicated by the extent to which the
standing of an applicant on the selection composite score corresponds to
his or her standing on a performance measure acceptable to the employer
as a criterion of job performance. As a general rule, the reliability
of the criterion sets an upper limit to the validity that could be ex-
pected for any empirical sample.

The criterion measures employed in controller selection research can
be divided into two categories, representing measures obtained during
initial training, while the individual is still in the status of student,
and post-training, on-the-job, which includes both developmental status
and full performance level (FPL) or journeyman status. Use of initial
training measures, such as pass-fail in the FAA Academy, is meaningful
in relation to the effect of the quality of student input on Academy
attrition, but is enhanced by evidence that training performance is also
predictive of later, on-the-job performance. Predictive validity of con-
troller selection implies prediction of both during-training and post-
training criteria.

Criterion measures at both levels are complex and confidence in the
selection program depends on their credibility. Chapters 9, 10 and 11
describe the training program (up to August, 1981) and the approach taken
by FAA to measure student performance. Chapter 9, by James O. Boone,
describes the training for Terminal and EnRoute students and the assess-
ment of student performance in these options. Chapter 10, by Evan W.
Pickrel, covers similar ground for the Flight Service Station (FSS)
students. Chapter 11, by Evan Pickrel and Jack Greener, provides a
discussion of a separate proficiency test used in the past to measure
student progress in the Terminal and EnRoute training programs, but which
is being adapted also to similar use in FSS training. The latter develop-
ment, as well as the development of an entire genre of paper-and-pencil
proficiency tests, for use both during and subsequent to training, is
reported by Joseph A. Tucker, in Chapter 12.

The conceptual as well as operational complexity of post-training
performance measures is vastly greater than that of during-training measures.
This is reflected in Jack Greener's report on post-training, on-the-job
measurement of controller performance, in Chapter 13, which describes and

evaluates measures that have been used as criteria in the selection
research reported in this book.

The final chapter of Part III (Chapter 14), also by Jack Greener,
summarizes the state-of-the-art in performance measurement, with parti-
cular emphasis on the air traffic controller job. This chapter addresses
in addition new measures that have either not been used, or investigated
only experimentally in controller selection research, but that may be
available in the automated system environment of the future. These include
computer-scored measures based both on system simulation and on sampling
of actual performance on the job.

                                        Jack M. Greener

154.

Chapter 9

THE FAA AIR TRAFFIC CONTROLLER TRAINING PROGRAM,
WITH EMPHASIS ON ASSESSMENT OF STUDENT PERFORMANCE

James O. Boone

EARLY DEVELOPMENTS--1947 to 1976

1947. Decentralized Training

Air traffic control specialist (ATCS) training began in Oklahoma
City in February 1947, at what is now called the Mike Monroney Aeronau-
tical Center. Six students started that training in a small classroom
in the rear of wooden barracks located southwest of the city at a con-
verted war-time military airfield. In the ensuing period, from February
1947 to March 1956, the ATCS training program expanded from its early
beginnings to become the major centralized ATCS training program in the
world. Three different programs were offered. International students
from Europe and the Near and Far East were administered a 23-week basic
ATCS training program, covering both EnRoute (air traffic between air-
port areas) and Terminal (air traffic in and around airports) airspace
control. Military flight facility officers (military airport tower
chiefs) received a 12-week course consisting of a general ATC orienta-
tion and an introduction to EnRoute and Terminal control and Flight
Service Station functions (facilities where flight plans are filed,
and weather information can be obtained). A third 2-week course was
taught for aviation executives, which covered the basic principles of
air traffic control. However, there was no centralized program for
civilian controllers. While the Civil Aeronautics Administration (CAA--
predecessor to the FAA, which replaced it pursuant to the 1958 Federal
Aviation Act) administered the Aeronautical Center operation, its own
air traffic controllers were not trained at the facility. They were
trained at the various ATC operational facilities, using training materials
developed by the Oklahoma City staff.

1956. First Centralized Programs

In 1955-56 the CAA became concerned over the lack of standardization
in ATCS training. This concern led to the establishment, in March 1956,
of a centralized ATCS training program at the Aeronautical Center. This
program consisted of an 8-week pass-fail course covering both EnRoute and
Terminal airspace control. The training was divided into two phases, an
academic phase and a laboratory phase. Few failed the academic portion;
however approximately 30% of the entrants failed the laboratory portion,
in which simulated problems were graded by instructors who trained and
assessed the students throughout the training. Failing candidates were
returned to their respective regions as soon as it was determined that they
did not possess sufficient potential to continue. At the termination of
training, the surviving students were assigned to field facilities by a
process in which regional representatives met with instructors and assign-
ments were based on demonstrated potential, available positions, and the

candidates' preferences. This program continued until June 1957. Inter-national students and military flight operation officers continued to be trained at the facility during this time.

1957. Decentralization

The mid-air collision of two commercial airliners over the Grand Canyon in 1956 resulted in a strong impetus to increase the number of ATCSs in the field (Holbrook, 1974). In July 1957, centralized training at the Aeronautical Center was closed and training was transferred to the field. The Aeronautical Center group continued to provide the field with training materials. Simultaneously with this action, the military trans-ferred its training to the newly established Biloxi, Mississippi ATCS training center. International students continued to be trained at the Aeronautical Center, and indeed have been continuously trained at the facility until the present. During the later 1950's and the 1960's the European countries and the more sophisticated Eastern countries developed their own training centers. At present most of the international ATCS training provided at the Oklahoma City facility is for the less well de-veloped "third world" countries.

1958. Return to Centralized Training - AOS Program, Separate Options

Beginning in January 1958, and continuing through September of that year, a 4-week centralized ATCS academic training program was implemented at the Aeronautical Center on a pass-fail basis. The course was based totally on an academic curriculum known as the 7-part Airways Operations Specialists (AOS) course. It covered both EnRoute and Terminal airspace control, and included topics such as communications, weather, and prin-ciples of flight. On the last day of the course a 7-part test, covering each of the topics, was administered and a passing score was required on five of the seven tests. Those who failed one or two topics were retested on the failed topics on the following Monday, and a passing score was re-quired on all previously failed topics. The program resulted in a 10% failure rate.

Increasing concern for standardization in ATC led to the development, in 1958, of a course on flight strip-writing at the Aeronautical Center. Flight strips are small strips of paper placed in removable holders and used by controllers to record with various symbols the history of each flight passing through their airspace. In October 1958, a 2-week pass-fail course in Flight Strip-Writing was added to the 4-week, 7-part AOS course. Because of differences in the airspace and strip-writing features of EnRoute and Terminal ATC, a separate laboratory and grading laboratory was established for each option. This was the first time that candidates were assigned, trained, and assessed at the Aeronautical Center in separate options. This process of early option assignment set a pattern that has continued until the present. With the addition of strip-writing, the Terminal failure rate remained, as before, at 10%, while EnRoute, which has a more complicated strip-writing process, had a 15% failure rate. The program remained in this status until December 1959. During this period,

mid-1958, the training program was moved into its present classroom facilities and was officially titled the FAA Academy.

1960. Centralization of FSS Training, Addition of a Laboratory Phase

In order to establish a pass/fail course, better founded on actual demonstrated ability to control aircraft, the Academy pass-fail program was modified in January 1960, to an 8-week program, containing a laboratory phase in addition to the 7-part AOS course and flight strip-writing course. Six laboratory problems were administered, and a passing score of 70 was required on four of the six problems. In January 1960, a centralized training program was initiated for the Flight Service Station segment of ATC. Failure rates for the pass-fail EnRoute and Terminal programs were 28% and 18%, respectively.

1962. Decentralization Again

From December 1962 through December 1965, a sharp curtailment in ATCS recruitment led to the elimination of centralized training at the Academy. New automation systems, the Automated Radar Terminal System (ARTS) and the Stored Program Alpha-numerics (SPAN) system, were purchased, with the expectation that a reduction in manpower would be achieved. These projections, however, did not take into account the sharp increases in air traffic that developed, and the reductions were not realized. During this period, specialized courses were provided at the Academy to journeyman controllers, first-line supervisors in air traffic, and regional office personnel. The Academy also continued to provide field training curriculum materials. To assist in developing course materials, a section of educational specialists was added to the staff. Assistance from educational specialists in course development has continued to the present.

1966. Resumption of Centralized Training, JTA-Based Curriculum, Technical Assessment.

The 1960's saw the development in general educational theory of a systems approach to educational programs, based on job task analysis (JTA). Since the air traffic control occupation had changed considerably, as a result of the automated systems, the agency embarked on the development of a new ATCS centralized training program based on job task analyses (Cobb, 1962; Trites, Miller, and Cobb, 1965;). In December 1966, the Academy began anew its centralized training, based on a JTA. The primary change that occurred, based on JTA, during this era was in the academic areas. During 1966-68, the pass-fail program consisted of academic topics only. At first, these were limited to the 7-part AOS course, with a slow evolution to the JTA-based curricula. New JTA-based academic curricula were completed in 1969 and replaced the previous 7-part AOS course. Academic evaluation consisted of a comprehensive phase test over the entire academic curriculum. When laboratory problems were introduced, the assessment process for laboratory problems was modified. In addition to the instructor's assessment on a 0 to 100 scale that had been previously used to score candidates, a count of various errors was used, termed a technical assessment.

157.

On the technical assessment, a student began the problem with 100 points
and various points were subtracted from the 100 points, according to the
type of error. The final laboratory problem grade consisted of an average
of the instructor assessment and the technical assessment. Laboratory
problems during this time were configured where the EnRoute and Terminal
sectors were adjacent and interactive. This program resulted in a 25%
EnRoute failure rate and a 15% failure rate for Terminal.

1971. Retention of centralization, Fair Employment Practices, Corson
Committee Recommendations

The 1960's was also a time of social introspection for the Federal
system. Concern about prejudice on the basis of sex and race in deci-
sions affecting employee's status resulted in the 1966 congressional
adoption of a set of Uniform Guidelines for Employee Selection (the
development of the guidelines is described in detail in the most recent
revision, 1978). The guidelines established criteria to be used in deter-
mining the validity of all governmental programs that resulted in decisions
affecting the employment status of any employee. While data had been
collected previously on various facets of the Academy pass-fail program, no
systematic data base had been established to demonstrate satisfaction of the
guidelines criteria. Further, the assessment process was based largely on
expert judgment by full performance level controllers. The Air Traffic
Controller Career Committee report (Corson, 1970) pointed out inadequacies
in the reliability and validity of the program. During this period, the
agency was also experiencing labor relations problems with The Professional
Air Traffic Controller Organization (PATCO), in the form of "sick-outs"
over various complaints, among which were training deficiencies. Be-
cause of these various considerations, beginning July 1971, centralized
training was continued, but pass-fail at the Academy was suspended until
a fully validated program could be established. From 1971 through June
1975, ATCSs who had been in field training for approximately 1 year were
brought to the Academy for qualifications training. They were trained
and assessed on their ability to operate simulated air traffic on sectors
taken from their home facility. The previously described assessment pro-
cedure was continued during this (no pass-fail) era; however, the scores
were simply returned with the ATCSs to their facilities, to be used pri-
marily for placement decisions.

1975. Development of Validated and Standardized Central Pass-Fail Training
Program

From 1970 through 1975, a series of studies was performed by the Air
Traffic Controller Career Committee (Corson, 1970), the Institute for Defense
Analysis (Henry, Karmrass, Orlansky, Rowan, String, and Reichenback, 1975),
and the Committee on Government Operations (1976), listing several recom-
mendations relating to ATCS training. In summary, the reports recommended
the development of a validated, standardized, and centralized pass-fail
training program, designed to screen as early as possible those persons
who lacked sufficient potential to become fully certified ATCSs. The
primary burden of early screening was to be focused on the Academy program,
in order to relieve the field facilities for more on-the-job type training.
Early Academy screening was to consist of both a nonradar and radar training

158.

phase. Further, the Systems Development Corporation (1972) completed
an updated JTA to be used as a basis for training curricula. To meet
the increasing demands in ATC, brought about by increasing air traffic,
further automation was also sought. Recognizing this need, Congress
enacted the 1970 Airport and Airway Development Act, beginning a trust
fund as a source of revenue to purchase automation equipment (Holbrook,
1974).

The FAA responded to these new requirements by forming a task group
to review and formulate a comprehensive program for ATCS staffing,
screening, and training. This group set in motion the establishment
of a fully validated, centralized Academy pass-fail training program
with a tracking component, to utilize a continuing data base to monitor
the quality and effectiveness of the program, and to maintain compliance
with the Uniform Guidelines for Employee Selection (FAA Task Force Report,
1975). This brings the historical review up to the present program, which
is discussed in detail below.

CURRENT ACADEMY TRAINING PROGRAM, SINCE 1976

Student Assessment - Academic Phase

In January 1976, the present pass-fail Academy program was implemented,
based on nonradar ATC. In July 1980, a new radar training facility, discussed
later, was opened and began training students on a non-pass-fail basis. The
structure of the present ATCS pass-fail Academy program is similar to the
preceding 1971-75 program; only the methods employed for student assess-
ment and the time of entrance for candidates have been changed. Following
a 1-week orientation, new selectees in the present program are sent directly
to the Academy. EnRoute training consists of two pass-fail Academy phases,
and the Terminal curriculum consists of two pass-fail Academy phases and one
non-pass-fail phase. (The Flight Service Station program which is not con-
cerned with the control of transiting aircraft, is not included in this
review.) Both EnRoute and Terminal training have an academic phase and a
laboratory phase. Terminal has an additional non-pass-fail Visual Flight
Rules (VFR - noninstrument flight) tower training phase. The academic
phase contains the same material for both EnRoute and Terminal activities,
while the laboratory phases conform to the specifics of EnRoute or Terminal
airspace control.

Block and Phase Tests. The academic phase consists of ten blocks of
instruction, listed in Table 1. Following each block of instruction, a block
test is administered to determine how well each candidate has mastered that
block. After all blocks have been covered a comprehensive phase test (CPT)
is administered to measure retention for each student across all the academic
material. Scores for block tests average around 95, and CPT scores average
around 92.

Nonradar Laboratory Instructor Evaluation. The laboratory phase is
the most important portion of the assessment process. Laboratory problems
are based on control of simulated nonradar air traffic. An instructor seated
behind the candidate maintains a log of the errors committed by the candidate

159.

Table 1

Subject Areas of Phase II

1. Air Traffic Service
2. Principles of Flight
3. Aircraft Types and Characteristics
4. Meteorology
5. Navigation
6. Federal Air Regulations (FARs)
7. Air Traffic Control Communications
8. Flight Assistance Service
9. Fundamentals of Radar
10. National Airspace Systems

on a structured worksheet (see Appendix 1 for a sample copy). Errors
are categorized as (1) separation errors, a violation of rules related
to minimum distances between aircraft or obstructions (2) coordination
and procedure errors, primarily violation of rules in interfacing with
adjacent airspace sectors or facilities and procedures that result in
excessive delays, and (3) other errors, a miscellaneous category which
includes strip-writing errors and incorrect phrasing of air traffic in-
structions to aircraft.

The frequency of errors in each category is transferred to an evalua-
tion form for calculation of the problem grade (see Appendix 2 for a sample
copy). A candidate begins the problem with 100 points. For each separa-
tion error, 15 points are subtracted, to a maximum of 50 deductible points;
coordination/procedure errors result in a loss of 4 points for each error,
to a maximum of 40 deductible points, and other errors are deducted at the
rate of 1/2 point each to a maximum of 10 deductible points. The score
received from the error counting process is termed the technical assess-
ment (TA). At the bottom of the evaluation form a rating scale from 40
to 100 is provided for an instructor assessment. The instructor assess-
ment for each problem is based on the instructor's evaluation of how well
the candidate demonstrated the potential to become a fully certified ATCS
on that problem. It is based on attributes not inherent in the error
counting process, such as selection of appropriate separation methods
and how the candidate structured the air traffic. This rating is termed
the instructor assessment (IA). The total score for the problem is the
arithmetic average of the TA and IA. A score of 70 is considered passing.
The IA is limited on the low end at 40 points to insure that no candidate
who scored 100 on the TA could have a failing problem total score due to
the IA. When the evaluation form is completed and the problem total score
is calculated, the results are reviewed with the candidate and the candi-
date and instructor sign the form.

The entire laboratory phase for both EnRoute and Terminal training
consists of six graded problems of increasing difficulty. (A large number
of practice problems are interspersed among the graded problems.) The
laboratory average is formed by weighting the first two problems 10% each
and the last four problems 20% each.

Controller Skills Test. The last measure of the Academy assessment
process is a standardized achievement test, consisting of 50 multiple
choice items requiring application of ATC principles to solve a variety
of ATC problems. The test was developed and validated using field facility
and military ATCSs as subjects, and is continually updated by placing un-
scored items in the test, for item analysis, which are used in later forms
of the test. It is titled the Controller Skills Test (CST). (Development
of the CST is reported in detail in Chapter 11.)

Calculation of Pass Fail Composite. The total composite used to determine
pass-fail is a weighted sum of (1) the averaged block test scores, weighted 2%,
(2) the comprehensive phase test score, weighted 8%, (3) the laboratory average,
weighted 65%, and (4) the CST score weighted 25%. To allow for the possibility

161.

of one unreliable laboratory problem, the best combination of five of
the six laboratory problems (the five that maximize the composite score)
are employed in calculating the total composite grade. The weight of
the dropped problem is distributed proportionately across the remaining
five problems in computing the composite score. A score of 70 or higher
is required to pass the Academy program.

ANALYSIS OF ACADEMY OUTPUT 1976-80, ATTRITION RATES IN ACADEMY AND ON THE JOB

The entire ATCS Academy program is monitored by the Avaiation Psycho-
logy Laboratory at the FAA Civil Aeromedical Institute (CAMI), using a
comprehensive program evaluation model. (Development of materials and
assessment instruments is a joint effort involving technical content experts
and a staff of educational specialists at the Academy, with occasional
input from CAMI psychologists.) The model covers four areas: (1) design
evaluation, (2) implementation evaluation, (3) formative evaluation, and
(4) summative evaluation. Design evaluation is an assessment of the com-
prehensive implementation plan; implementation evaluation is a determina-
tion that the plan is implemented completely and accurately according to
prescription; formative evaluation is a continual monitoring of the program
to keep the process reliable, stable, and on track; and summative evaluation
monitors the product of the training program. The design evaluation relies
on the task, knowledge, and skills analysis and on the documents in the
implementation plan. The implementation evaluation makes use of data from
frequent status studies. Formative and summative evaluations make use of
statistical reports and mathematical modeling, primarily linear regression
models, to monitor the process and products of the program and to estimate
and determine the impact of changes made to the program (Boone, 1982).

Academy Pass Rates, 1976-1980; Terminal and EnRoute Options

Several routine formative evaluation reports are used as information
in monitoring the program; however, this discussion is limited to summary
data. Figure 1 contains a summary of the formative data collected on pass-
fail rates. The summary covers the period from January 1976 through December
1980. Pass rates, shown in Figure 1 indicate that the Terminal program had
a success rate approximately 8% higher than the EnRoute program. Since
candidates were assigned to an option (Terminal or EnRoute) based on
available job openings rather than on some measured attribute, it appears
that the Terminal program is the easier of the two options. The same
pattern holds when the pass rates are distributed across the years of
1976-80, as shown in Figure 2. It should also be noted in Figure 2, that
during 1976-77, while the program was being refined and more weight was
placed on the laboratory scores in computing the composite score, a decline
in pass rates occurred. Since then, the pass rates have been somewhat more
stable.

Tables 2 and 3 and Figures 3 and 4 show pass rates by sex and minority
group. The general trend is that nonminority men had the highest pass rates;

162.

ACADEMY PASS/FAIL RATES
1/76 – 12/80
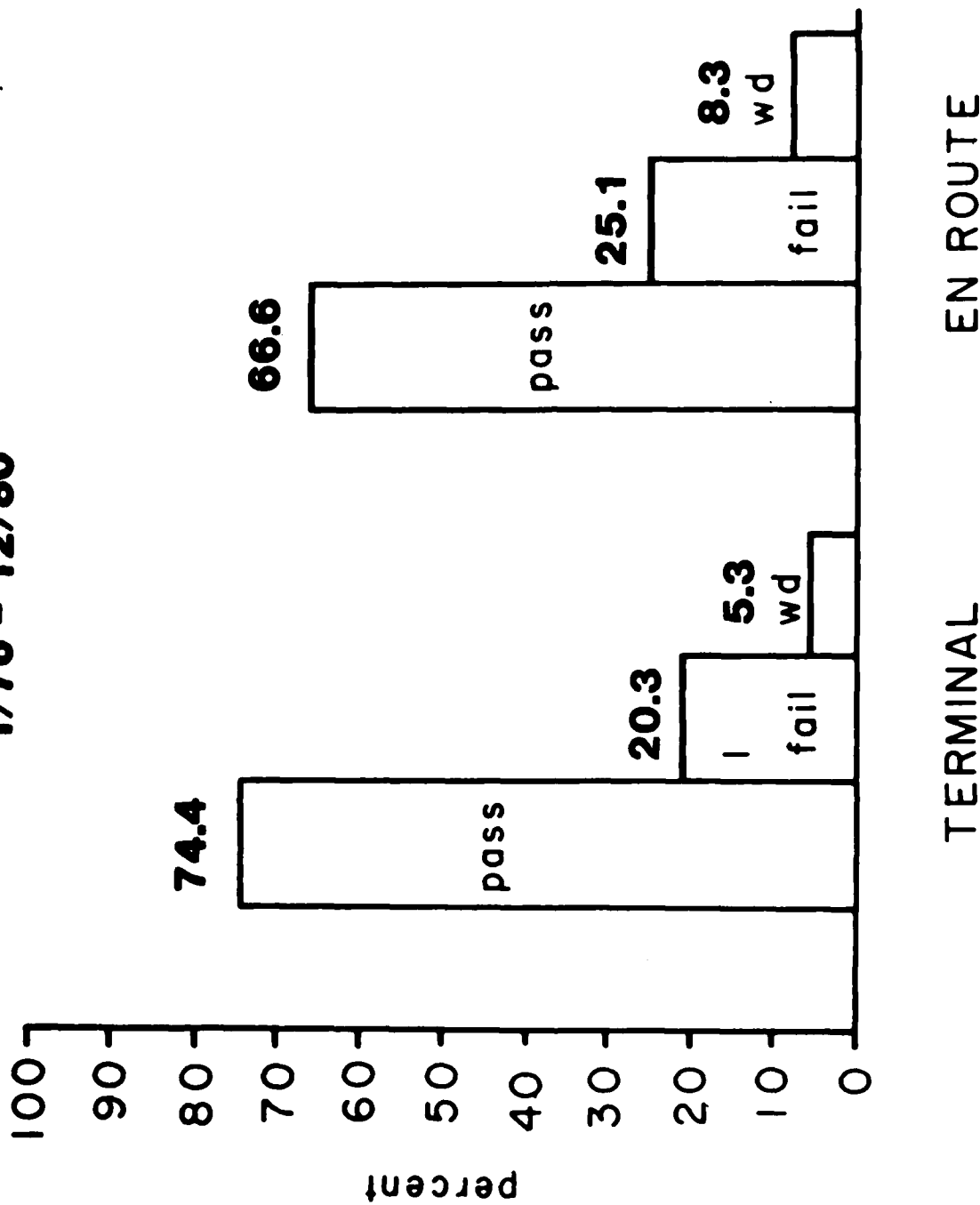
Figure 1. Academy pass-fail rates.

163.
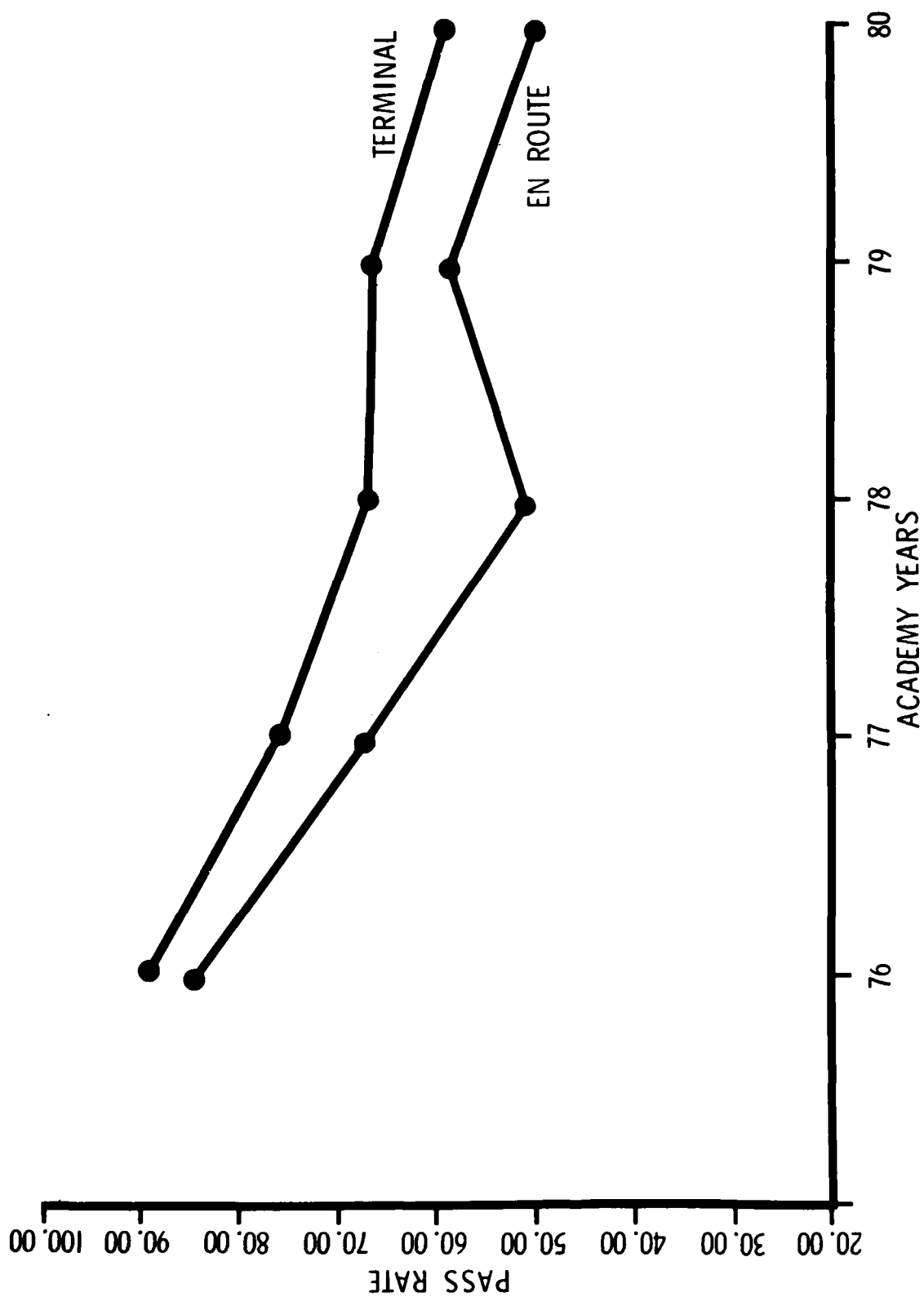
Figure 2. Terminal and EnRoute Academy pass-fail rates by year, 1976-1980.

164.

Table 2

Terminal Nonradar Lab Phase Mean Scores

1976 Through 1980 Entries

	Total	Women	Men	Minority	Nonminority
No. of Entries	3962	669	3293	526	340
Pass Rates	74.41%	65.62%	76.19%	55.32%	77.36%
Composite	75.28%	73.05%	75.72%	70.16%	76.02%
Block Tests	95.39%	95.36%	95.39%	94.26%	95.54%
Phase Test	91.65%	91.02%	91.77%	89.59%	91.95%
Controller Skills Test	76.42%	74.34%	76.83%	71.05%	71.19%
Lab Average	72.83%	70.18%	73.35%	67.43%	73.61%
Technical Assessment (TA)	65.35%	61.44%	61.12%	57.68%	66.51%
Instructor Assessment (TA)	76.87%	75.24%	77.20%	71.74%	77.65%
IA – TA	11.52%	13.80%	11.08%	14.06%	11.14%

Table 3

Enroute Nonradar Lab Phase Mean Scores

1976 Through 1980 Entires

	Total	Women	Men	Minority	Nonminority
No. of Entries	4129	598	3530	434	3647
Pass Rates	66.55%	57.69%	68.05%	58.06%	67.59%
Composite	73.74%	71.48%	74.11%	71.08%	74.08%
Block Tests	95.67%	95.42%	95.71%	94.71%	95.79%
Phase Test	93.73%	93.09%	93.84%	92.56%	93.86%
Controller Skills Test	77.01%	74.17%	77.47%	72.65%	77.55%
Lab Average	69.77%	67.53%	70.13%	67.22%	70.09%
Technical Assessment (TA)	61.32%	57.81%	61.88%	58.94%	61.64%
Instructor Assessment (IA)	75.90%	74.13%	76.19%	73.76%	76.15%
IA - TA	14.58%	16.32%	14.31%	14.82%	14.51%
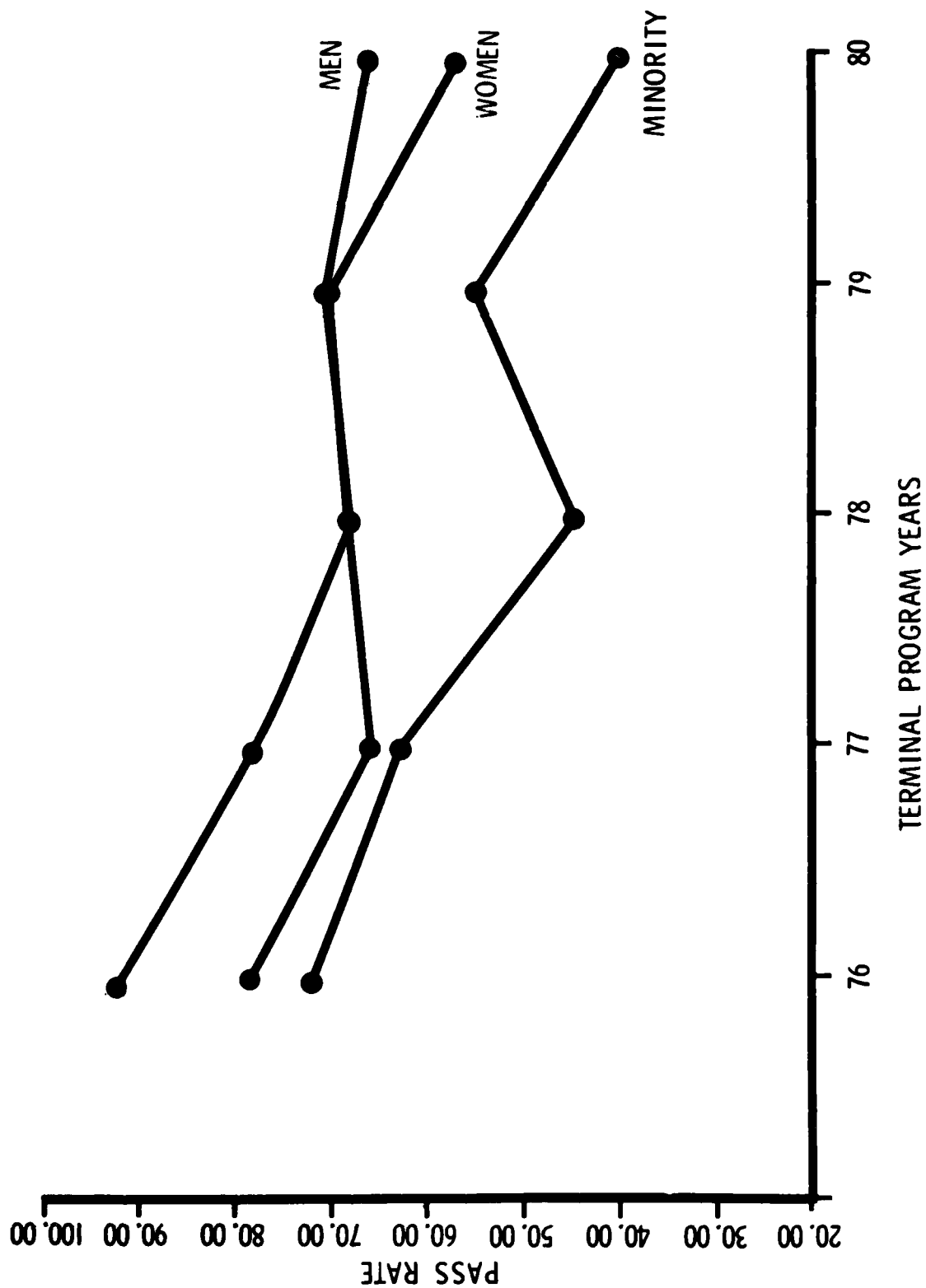
Figure 3. Terminal program pass-fail rates for men, women, and minorities, by year, 1976-1980.

167.
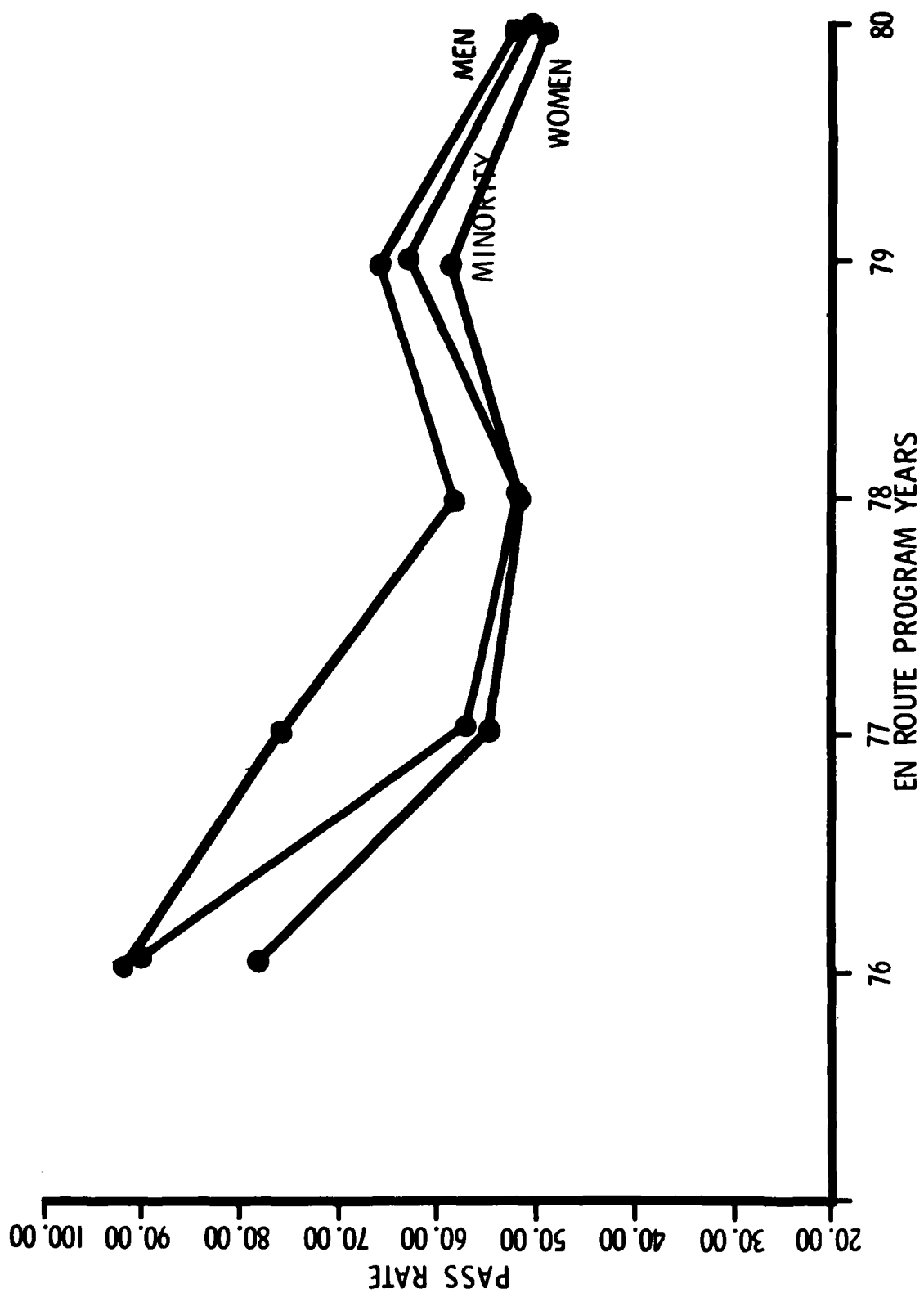
Figure 4. EnRoute program pass-fail rates for men, women, and minorities, by year, 1976-1980.

168.

nonminority women had the next highest pass rates; and minorities had the lowest pass rates. In the years 1978-79, in the Terminal option, the pass rates for nominority women were nearly equivalent to those for nonminority men. This was an effect of a special recruitment program to increase the number of women and minorities in ATC (Boone, 1978). As Figure 3 demonstrates, the primary effect of the program was on nominority women. For a variety of reasons, the program has been discontinued.

Tables 2 and 3 also show the average scores across all the measures included in the computation of the total composite score, which has been used to determine pass-fail. Mean scores on the academic phase block tests and comprehensive phase test have been quite high, at 95 and 92 respectively. The applied measures, such as CST scores and laboratory scores have been much lower, with means of 75 and 71 respectively. The difference between the technical assessment and instructor assessment on laboratory problems shows what appears to be a slight halo effect in favor of women and minorities.

Effects of Experience Variables on Pass Rates

The effects on pass rates of having past ATC experience, related aviation experience (e.g., pilot experience), or no related experience are illustrated in Figures 5 and 6, for Terminal and Enroute trainees respectively. Across the years, from 1976-80, the highest passing rate was obtained by persons having prior ATC experience. Specifically, this effect reflects high pass rates for students who entered the Academy with prior Instrument Flight Rules (IFR) radar ATC experience; candidates who entered with prior Visual Flight Rules (VFR) ATC experience had pass rates similar to those of the "related aviation experience" group, which was second highest. While candidates who entered with prior aviation experience passed the Academy training with higher rates, CAMI studies have shown that they attrited at a higher rate than the "no experience" group (the group having the lowest Academy pass rates) during subsequent on-the-job training in the field; as a result, the overall attrition rate was approximately equal. The lowest attrition rates at the Academy and in on-the-job training occurred among candidates who entered the Academy with prior IFR ATC experience. These results are consistent with earlier studies (Trites and Cobb, 1964; Cobb, Nelson, and Mathews, 1974)

Post-Academy Attrition by Sex and Minority Status

Post-Academy attrition rates by sex and minority status are listed in Table 4. Summative evaluation data used to monitor the quality of the ATCS output from the Academy are collected for candidates beginning around 1.5 to 2 years after graduation. Approximately 3.5 to 4 years are usually required to become a full performance level FAA certified ATCS. In the discussions of Tables 2 and 3, it was noted that the Academy pass rates for women and minorities were lower than for nonminority men. In Table 4 the post-Academy attrition rates are shown for these groups and they were almost precisely equal, with the exception of nonminority women. However, when attrition was analyzed separately for Post-Academy failure and

169.
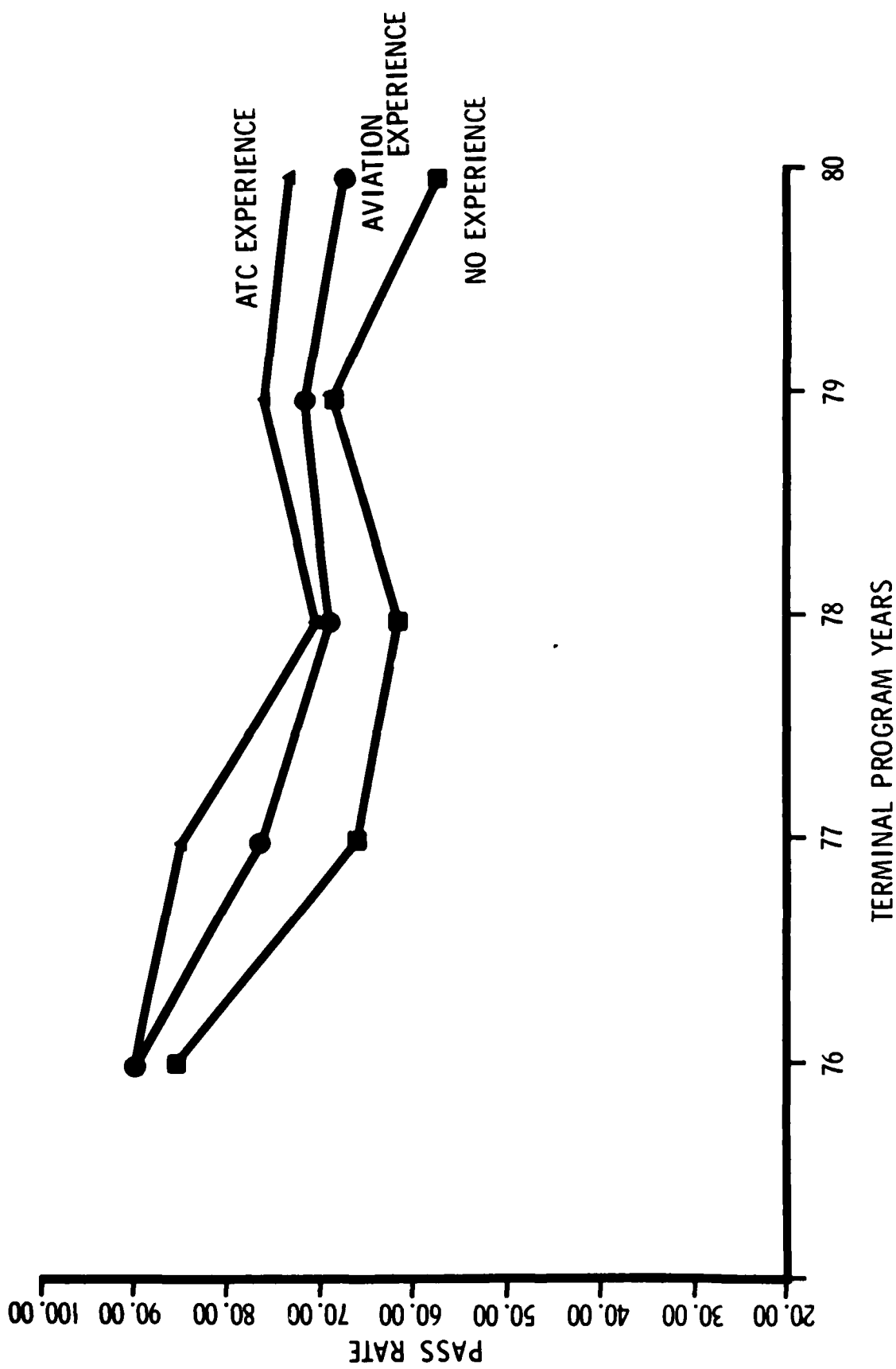
Figure 5. Terminal program pass-fail rates for 3 categories of prior experience, by year, 1976-1980.
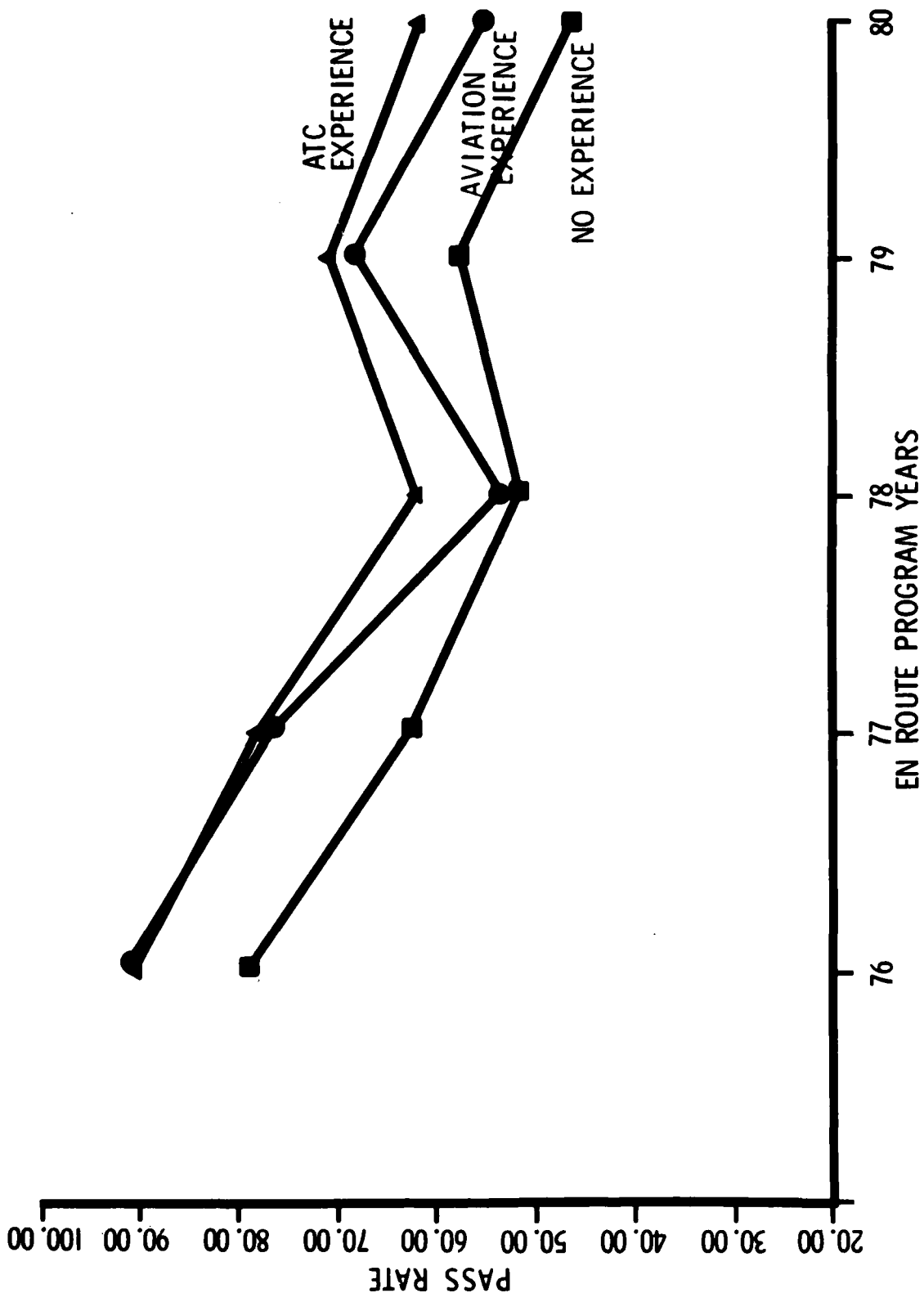
Figure 6. EnRoute program pass-fail rates for 3 categories of prior
experience, by year, 1976-1980.

Table 4

Post-Academy Attrition Rates

May 1977-January 1978 Entries

	Men	Women	Minority	Non-Minority	Total
Post-Academy Active Rate	85.1%	79.2%	84.4%	84.5%	84.5%
Loss Rate	14.9%	20.8%	15.6%	15.5%	15.4%
Failing	6.0%	6.2%	6.7%	6.0%	6.0%
Other Reasons	8.9%	14.6%	8.9%	9.5%	9.4%

172.

other (usually personal) reasons, the post-Academy failure rates were almost equal for all groups. The higher attrition rate for women was attributed mainly to "other" reasons, which according to CAMI studies, include pregnancy, getting married, unwillingness to be career mobile, and the like (Mathews, Collins, and Cobb, 1974). This evidence demonstrates that while disparities exist in Academy pass rates, according to sex and minority status, the Academy produces a uniform product without regard to sex or minority status. CAMI studies across time have also shown that while the field training failure rate is strongly affected by Academy training, as reflected by the pass rates, attrition for other reasons remains stubbornly around 10% (Trites, 1961; Cobb, Mathews, and Nelson 1972). Since attrition for other reasons is a result of factors outside the scope of FAA training, most efforts designed to reduce field attrition have been aimed at improvement of skills.

Analysis of Field Supervisors' Global Ratings; Support for Centralized Training, and Early Screening of Ineffective Controller Candidates

One part of the summative evaluation is a supervisor rating of each candidate in several areas of performance and an overall global rating of proficiency in ATC. The global rating scale is a judgment by the supervisor as to whether the candidate is inadequate, marginal, below average, average, very good, excellent, or outstanding. Table 5 shows how these ratings distribute across the Academy laboratory scores. Since only those who score above 70 are passed to field training, laboratory scores for the candidates start at 70. In the column labeled Inadequate, Marginal, Below Average, the percentage of graduates rated in these categories dropped consistently from 28.5% for laboratory scores between 70-74 to 0.0% for laboratory scores between 90-100. All the columns followed a similar pattern except the mid-level Very Good column, which remained essentially constant over the range of laboratory scores. In the laboratory score range of 85-89, only 2% were rated as below average while 96% were rated as very good to outstanding. In the 90-100 laboratory score range, 100% of the graduates were rated as very good to outstanding. Reading across columns in the laboratory score range of 70-74, the Inadequate, Marginal, Below Average column accounted for 29% of the graduates, while the Excellent-Outstanding column accounted for 31%. The high percentages of low scores in the Very Good, Excellent, and Outstanding categories are believed to illustrate the commonly observed "halo effect" in supervisor ratings. However, the remaining laboratory score range categories across columns do not follow the halo pattern. The conclusion drawn from this table is that a high relationship exists between Academy laboratory scores and supervisor rating of proficiency in on-the-job performance 1.5 to 2 years after Academy graduation.

What is the bottom line benefit to the FAA of Academy training and screening? Table 6 presents a comparison of training failure attrition rates in the Academy and in the field for two time periods: (1) the 1971-75 era, during which the Academy was not organized on a pass-fail basis, and (2) the period 1976-80, during which the present pass-fail system has been in practice. For the earlier period, no training failure

Table 5

Performance in Field Training 1.5 to 2 Years after Academy
May 1977–January 1978 EnRoute Entries

Percentage of those in lab score category
who were rated by field supervisor as –

Academy Lab Phase Score Category	Inadequate Marginal Below Avg.	Avg.	Very Good	Excellent Out- Standing
70–74	28.5%	17.1%	23.6%	30.9%
75–79	15.1%	18.5%	22.7%	43.7%
80–84	11.4%	14.4%	18.9%	55.3%
85–89	2.0%	2.0%	22.4%	73.5%
90–100	0.0%	0.0%	22.2%	77.7%

Table 6

Bottom Line on Academy Screening

| | Attrition | |
Time Period	Academy	Field
Prior to Academy pass/fail	---	38%
Post Academy pass/fail	30%	7%

occurred at the Academy and the field attrition rate was 38%. The average length of time that these attritees worked for the FAA was 2 years (FAA Task Force Report, 1975). The primary attrition in the field occurred during radar (IFR) ATC training. During the 1976-80 period, a rate of 30% attrition occurred at the Academy with an average employee tenure before attrition of about 15 weeks. The current 7% attrition rate attributed to field training failures occurs after an employee tenure of 2 years. The estimated cost avoidance due to early Academy screening, according to figures provided by the Office of Management and Budget (OMB), is approximately $14 million annually, a substantial savings for an effective screening program. This information must be seasoned with the understanding that these estimates speak only to the <u>effectiveness</u> of the screening, not to the efficiency of Academy screening. Indeed, while the information taught at the Academy is essential for all students, it is conceivable that a highly tailored and streamlined computer-based instruction (CBI) program using simulation, could be more efficiently employed at field facilities with the same success in screening without incurring the additional millions of dollars in per diem and travel expense required to bring the trainees to the Academy. The feasibility of such a program operated under the auspices of the Academy is presently being evaluated.

DESCRIPTION OF THE ACADEMY RADAR TRAINING FACILITY

As previously stated, the Academy completed a Radar Training Facility (RTF) in 1980, which was designed to provide some radar training and at the same time to screen out those who failed to demonstrate sufficient potential to advance to radar ATC. The primary objective of the RTF is to duplicate closely the specialized operational environment existing at automated Terminal and EnRoute facilities, with the capability to synthesize and present a wide variety of air traffic control situations. The plan for the RTF is to utilize situations based on a reference data base created through scenario programs with a full range of control necessary to establish realistic simulation of actual aircraft traffic under a variety of conditions.

To accomplish this objective, four independent laboratories are utilized. Figure 7 describes how the laboratories are configured.

Positions

There are Trainee positions and Supervisory and Support positions/stations corresponding to each radar training sector. At each "position," the operating personnel have input/output (I/O) equipment access and in addition, are equipped with voice communications for monitoring, instructing, and supervisory functions.

Trainee Position

1. <u>Radar Control Position</u> (R). The R controller positions (six in each lab) have a display console, (PVD) for EnRoute and (DEDS) for Terminal. They have associated voice communications. The displays and voice communications are similar to those at field facilities. Displays include maps,

Figure 7. Laboratory configuration.

P - Pilot Console
G - Ghost Console
R - Radar Training Console
M - Manual Controller Trainee Position
    En Route "D" Controller or
    Terminal "HO" Controller

MI - Master Instructor Position
PS - Pilot Supervisor Station
I - Instructor Station
FSP - Flight Strip Printer

weather, aircraft position symbols, alphanumeric readouts, and other digital and symbolic data.

2. <u>Nonradar Controller Position (HO/D)</u>. The D controller for EnRoute and the HO position for Terminal (six in each lab) have the capability of making and accepting handoffs. This position also permits training for manual or nonradar control by using flight progress strips generated by the flight strip printers.

3. <u>Pilot Position (P)</u>. Three P positions are associated with each sector (18 in each lab). These positions are in a separate room. Each position operator performs at a console with a tabular display and keyboard for data entry with associated voice communications. These operators simulate aircraft pilots during exercises by actual response to ATC clearances/instructions.

4. <u>Ghost Position (G)</u>. This position is associated with each R and/or HO/D position. There are six G positions in each lab. The position console and display are identical to those of the P position. The G position operator adds realism to the exercise by performing related functions of adjacent centers, terminals, flight service stations, and position/sectors. Functions include initiating handoffs, accepting handoffs, and general ghosting functions of other facilities/sectors.

Supervisory and Support Positions/Stations

1. <u>Instructor Station (I)</u>. An instructor station is provided at each sector (six in each lab). The instructor has voice communication with each student and monitors the overall exercise from behind the trainee positions.

2. <u>Pilot Supervisor Station (PS)</u>. This position (one in each pilot room) has voice communications for supervision, monitoring, and instructional operation of P positions as well as for coordinating activities with the master instructor station and the system monitor position.

3. <u>Master Instructor Station (MI)</u>. This position (one in each lab) controls the exercises within the laboratory. The position has a tabular display, a data entry keyboard, and associated voice communications with each trainee and with each operator of G,I, and P positions in the lab. The MI permits the setting of clock time, starting, monitoring, freezing, backing up, replay, and restarting of the exercises. This position also provides for data recording and analysis of the exercises.

4. <u>System Monitor Position (SM)</u>. One SM position is provided for each lab. This position has voice communications with two MI positions and two PS positions, and permits computer operation and operational and maintenance monitoring.

Figure 8 describes the system configuration for operation of the positions and stations in each laboratory. The training sectors are

178.

SYSTEM MONITOR
CONSOLE

PDP-11/34 PROCESSOR
● 2 OR 3 PER LABORATORY

SECTOR STUDENT CONSOLE
● 6 POSITION PER
LABORATORY

18 PILOT & 6 GHOST POSITIONS
PER LABORATORY
● 3 PILOT & 1 GHOST
POSITIONS FOR EACH
STUDENT CONSOLE

DISK   DISK

PDP-11/60
PROCESSOR

BUS SWITCH

MASTER INSTRUCTOR
CONSOLE

HIGH SPEED
PRINTER

FLIGHT STRIP
PRINTER
● 3 FLIGHT STRIP
PRINTERS (1 PER
STUDENT PAIR)

MAG
TAPE
UNIT

Figure 8.   Computer system configuration.

179.

controlled by a Digital Equipment Corporation (DEC) PDP 11/60 computer with a PDP 11/34 computer serving as an interface between the PDP 11/60 and the operating positions.

The training process involves three sequential systems of operation: (1) SCENARIO GENERATION --/ (2) REAL-TIME --/ (3) PERFORMANCE MEASUREMENT. Scenario generation, illustrated in Figure 9, is the non-real-time process of building exercises and evaluation problems for the system. Aircraft characteristics, flight plans, and other essential information of this type are stored in the Universal Data Files (UDF). Exercises are built by first selectively retrieving intermediate files and then creating other intermediate data files from the universal data base through the scenario management program.

The real-time component, illustrated in Figure 10, utilizes the scenario management files to generate the actual radar simulation exercise. The real-time component drives the display at the radar position. Aircraft movement is controlled through the P and G positions according to the instructions that the operators of those positions receive from the controller trainee or, in some cases, from a scenario prompt which appears on the cathode-ray tube (CRT) at the P or G positions. During the operation of a real-time training exercise, all actions taken during the exercise are recorded.

Student Assessment in Radar Training

At completion of an exercise, the computer analyzes the recorded actions to determine violations of separation standards and to quantify other pertinent performance information, such as delay times, in order to evaluate student performance and demonstrated ability to move air traffic "safely and expeditiously." The process of student performance measurement is illustrated in Figure 11.

Table 7 presents a list of the computer-derived measures employed in the valuation of student performance on problems. Studies performed by CAMI (Boone, 1980) indicate that the computer-derived measures show promise for application in the assessment of student performance, but at present software limitations and a need for further research preclude their use.

The curriculum design and assessment process for radar follows closely the design and process in the nonradar training phase. There are blocks of academic instruction, with block tests and a comprehensive phase test weighted 2% and 8% respectively, in the composite score, a laboratory phase weighted 65% in the composite, and a CST, weighted 25%, which is the same process as in nonradar. Laboratory problems are scored over-the-shoulder by an instructor on the same forms described in nonradar training. The basic difference between the radar and nonradar laboratory involves the radar screen, where students visually see the aircraft and associated ATC information related to each aircraft. The separation standards in the radar phase allow aircraft to fly closer together, and this increases the number

180.
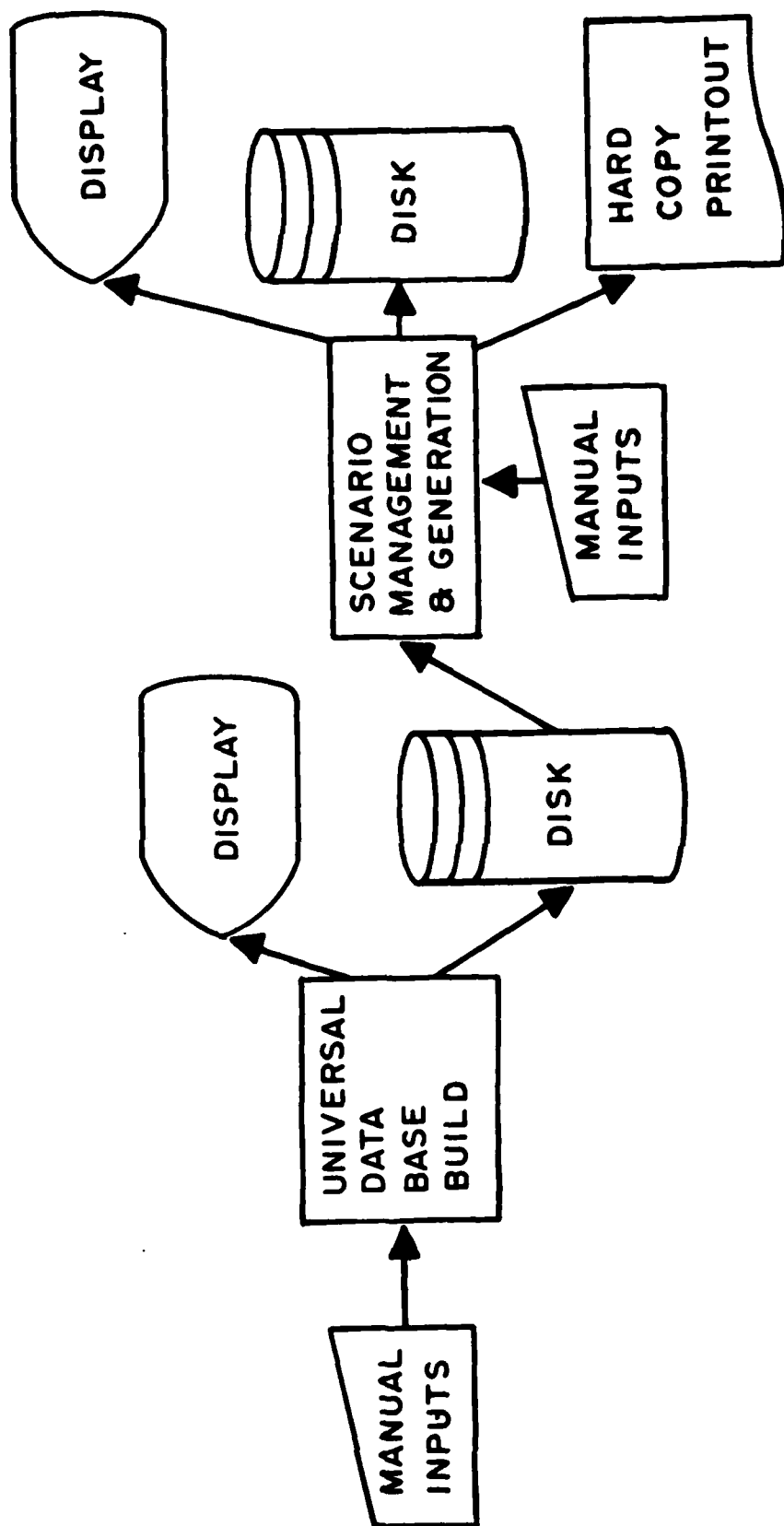
Figure 9. Components of scenario generation.
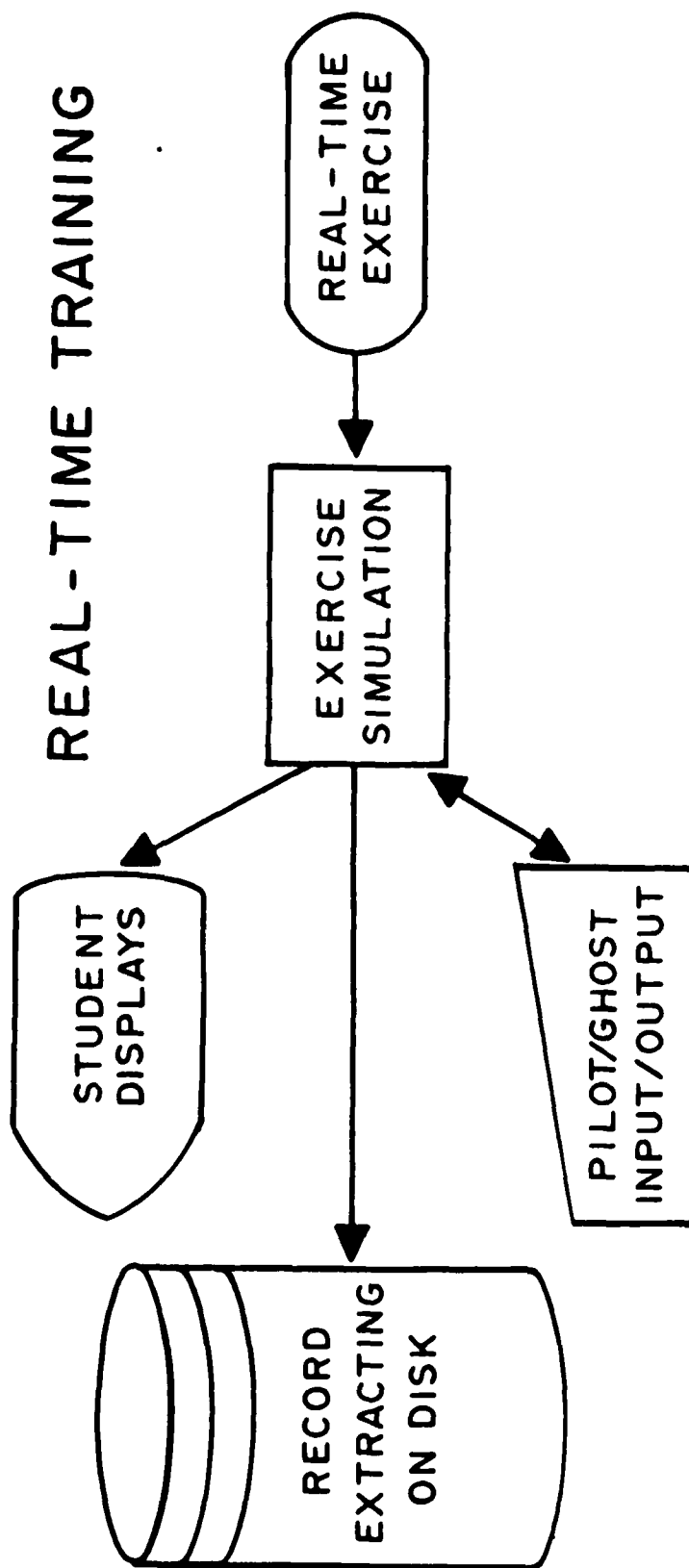
Figure 10. Components of the real-time training system.
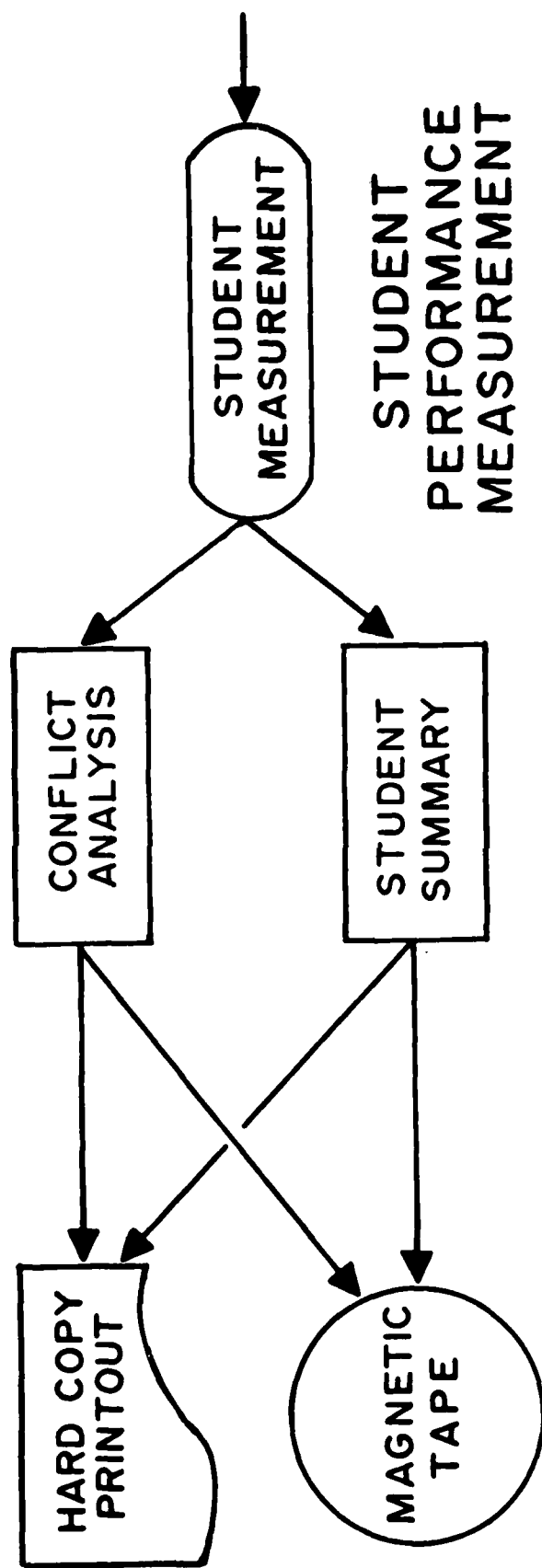
Figure 11. Components of the student performance measurement.

Table 7

Computer-Derived Measures and Their Corresponding
Reference Numbers Employed in the NAFEC Study

1. Conflicts (5-mile separation)
2. Conflicts (3-mile separation)
3. No. Start Point Delays
4. Start Point Delay Time
5. No. Turn and Hold Delays (turns longer than 100 seconds)
6. Turn and Hold Delay Time
7. Aircraft Time-in-System
8. No. Aircraft Handled
9. No. Completed Flights (transfers to 130.5 must be from ghost position)
10. No. EnRoute Departures (Code 2)
11. No. Terminal Arrivals (Code 3)
12. No. Terminal Departures (Code 4)
13. No. Air-to-Ground Contracts (subject only)
14. Air-to-Ground Communications Time
15. No. Altitude Changes (pilot keyboard messages)
16. No. Heading Changes (pilot keyboard messages)
17. No. Speed Changes (pilot keyboard messages)
18. No. of Handoffs From Feeder Position to Subject
19. Handoff Delay Time
20. No. Beacon Re-Idents

184.

of aircraft that the student controls in the radar setting. The radar
program is now undergoing validation studies and is not yet in a pass-
fail mode; as a result, quantitative information is not available for this
program. However, early indications show a failure rate of approximately
10% in radar training. When the program moves to a pass-fail basis, a
further reduction is expected in post Academy field training attrition.

THE PATCO STRIKE; EFFECTS ON THE SYSTEM; SYSTEM RECOVERY

In August 1981, an unprecedented strike of government employees was
staged by members of the Professional Air Traffic Controller Organization
(PATCO). Approximately 12,000 of the 18,000 ATCS workforce participated
in the strike. Emphasizing the illegality of the strike, President
Ronald Reagan first ordered the strikers to return to work within 48 hours
and subsequently ordered the discharge of those who did not return. At
the end of August 1981, the ATCS workforce consisted of approximately
6,000 controllers. Within a few weeks a recovery effort was in progress
to rebuild the system.

While no changes occurred at the Academy in the curriculum of the
pass-fail program, changes in the planned and actual Academy operations
were necessary. First, the RTF screening program which was scheduled
to occur immediately following the nonradar phase, was moved to phase X
about 1.5 to 2 years into field training, and just prior to radar training
in the field. Successful completion of the RTF is required for advancement
to radar ATC. Secondly, the FAA considered the Academy a major pivot point
in rebuilding the ATC system. The Academy had previously trained and screened
approximately 1,800 candidates per year. The recovery program required that
number to be increased to approximately 6,000 per year for 2 years, September
1981 through December 1983, while maintaining the same quality product, in
order to keep field attrition at a minimum. Recently retired ATCSs and
those medically disqualified to operate live air traffic were recruited,
screened, and hired on contract by the University of Oklahoma to perform
as Academy instructors, and a 3-shift operation began in October 1981. On
4-week cycles, 288 candidates in EnRoute, and on 5-week cycles, 216 candi-
dates in Terminal began training. As of January 1982, four EnRoute inputs
and one Terminal input have entered the Academy. Two of the EnRoute inputs
have completed training. At present the Academy recovery effort appears
successful; however the next two years most certainly will be challenging
years for the FAA Academy.

SUMMARY

The history of centralized ATCS training reveals an evolutionary
process of development leading to the present Academy program. Centralized
training at the Academy has been a cyclic process reverting back and forth
from centralization of training at the Academy to decentralized training at
field facilities. The same process is true with regard to pass-fail at the
Academy. Lack of standardization in field training at field facilities and
costly high attrition rates too far into training seem to have pressed the
pendulum toward centralized pass-fail training. However, lack of evidence

185.

concerning program reliability and validity have on occasion led to various attacks within and outside the agency, and these have tended to push the pendulum back toward decentralized field training, with pass-fail more firmly based on long term on-the-job training. The present program, which has been operative since 1976, has a strong data support system. This has enabled defense of the program, and it has survived and functioned success-fully up to the present. It is expected that the addition of the sophisti-cated Radar Training Facility described earlier will enhance future Academy success. The FAA continues to evaluate new methods to improve the effective-ness and efficiency of its Academy program in areas such as computer-based instruction. The recovery from large ATCS losses, due to the August 1981 strike, should prove to be the most strenuous challenge and test of Academy ATCS training and screening.

NOTE

The history of the air traffic controller Academy was obtained primarily from interviews with Benjamin Demps, Edwin Harris, Morris Friloux, Fred Fairweather, and Charles Hough whose combined experience as students, instructors, and administrators of the Academy span almost 30 years. The data presented on the Academy program from 1976-80 were collected and stored on computers by Drs. Mary Lewis, now with PPG Industries, James Boone, and Allan VanDeventer. The latter two summarized the data into the figures and tables presented in this report.

LABORATORY WORKSHEET   DATE 4/1/82

STUDENT NAME John Doe   PROBLEM NUMBER 15-1   SECTOR POSITION D2

| SEPARATION | 1 | ② | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | TOTALS 2 |

| PROCEDURES & COORDINATION | 1 | 2 | 3 | ④ | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 4 |

| OTHERS | | 1 | 2 | 3 | ④ | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 4 |

NO	A/C ID	ERROR	DESCRIPTION OF ERROR
1	NW344	P	O/R X 17NW MLC ↧60 SB ↧50
2	N443V	S	AA12 HPAS
3	N443V	O	PH DIDN'T USE SINGLE DIGITS
4	AA12	P	RESTRICTION UNREALISTIC (EXCESSIVE)
5	EA218	P	DIDN'T ISSUE REQ. ALT. WHEN AVAILABLE
6	UA538	O	S/M NO ALTITUDE CHECK
7	UA538	O	S/M NO PILOT EST. RECORDED.
8	UA538	S	VIOLATED ZKC AIRSPACE
9	N9WT	P	DELAY @ MLC
10	N9WT	O	DIDN'T REVISE EAC

STUDENT SIGNATURE John Doe   INSTRUCTOR Joe Jones

AC Form 3120-87 (12-76)

Appendix 1.   Laboratory worksheet.

187.

NONRADAR LAB EVALUATION FORM

Name: John Doe   Option: EN ROUTE   Phase: III   Date: 4/1/82   Problem: 15-1   Sector: TVL

SEPARATION 50%

Errors	0	1	2	3	4+
Weighted Score	50	35	(20)	5	0

20

COORDINATION/PROCEDURE 40%

Errors	0	1	2	3	4	5	6	7	8	9	10+
Weighted Score	40	36	32	28	(24)	20	16	12	8	4	0

24

OTHER 10%

Errors	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20+
Weighted Score	10	9.5	9	8.5	(8)	7.5	7	6.5	6	5.5	5	4.5	4	3.5	3	2.5	2	1.5	1	.5	0

8

TOTAL 52.0 Problem Avg.

= 86.0 Instructor Rating

Total 138

÷2

69.0 Lab Average

Ineffective ———————————————— Very effective

40 45 50 55 60 65 70 72 75 78 80 82 84 85 (86) 88 90 92 95 98 100

Yes / No

☑ ☐ My performance on this problem has been reviewed with me by the instructor.

☑ ☐ Questions I have asked regarding my performance have been adequately answered.

Archie Ross
Developmental Specialist Signature

Joe Jones
Instructor Signature

AC Form 3120-95 (5-77)

Appendix 2. Nonradar lab evaluation form.

Chapter 10

ASSESSMENT OF FLIGHT SERVICE STATION STUDENT PERFORMANCE

Evan W. Pickrel

OVERVIEW

A new Flight Service Station (FSS) pass-fail training program was implemented in 1978. It was designed to provide previously screened candidates with a training and evaluation curriculum to ensure that the great majority of them would achieve readiness for journeyman assignments, and at the same time, to eliminate the few whose performance in training indicates a high probability of failure on the job. This training program involves approximately 4 months at the FAA Academy, after which the student is assigned to a field facility to receive approximately 6 months of developmental training, and finally checkout as a Full Performance Level (FPL) specialist. This 10-month program is rapid, compared to the training for the Terminal and EnRoute options which may extend for 2 to 5 years before a student reaches the FPL level. (A discussion of early developments in air traffic controller specialist training appears in Chapter 9).

Training in each of the ATC options is keyed to an improved screening system that extends beyond initial selection testing and incorporates measures of the candidate's performance during training. In the FSS training program, pass-fail evaluation depends on evaluation in Phases II and III of the training program. When a student completes a phase in a satisfactory manner, he or she is advanced to the next phase. The calculus of pass-fail does not include Phase I, which involves 2 weeks of indoctrination on agencies, regions, and facilities, at a field facility. FSS Phase I is comparable to Phase I of the Terminal and EnRoute programs.

Phase II consists of 4 weeks of classroom training at the FAA Academy, and is almost identical to Phase II for the Terminal and EnRoute programs, but with more emphasis on weather, flight assistance service, and navigational aids. Training in this phase usually results in the achievement of high performance levels, and failures seldom occur. This is the initial ATC Academy pass-fail point; scores earned in Phase II are used to determine advancement to Phase III, but are not included in the composite for Phase III.

Phase III involves approximately 11 weeks of laboratory training and evaluation at the FAA Academy on the following duties: Preflight and In-flight Operations, emergency services to Aircraft, Weather Observer, Broadcast, Teletype, and Flight Data. The first three involve functions which, if not fulfilled, could have potentially catastrophic results and impact the safety of the air traffic system. These are also the more complex operational activities, requiring some performance of most of the duties of the other positions, and thus are at the top of the FSS position hierarchy. Final School Grade (FSG) at the FAA Academy is derived from a weighted composite of grades in Phases II and III, and is a determiner of career progression to Phase IV.

189.
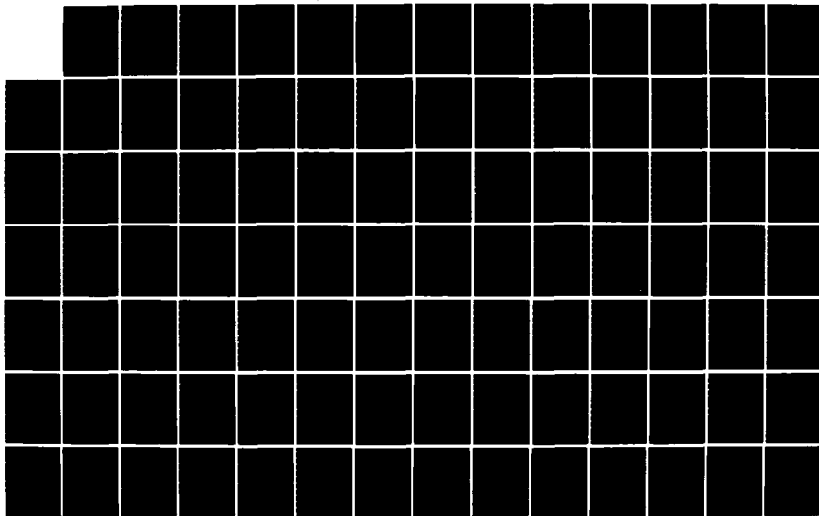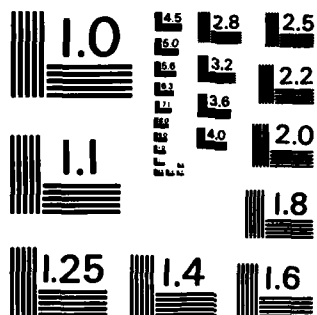
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Phase IV, Position Qualification and Facility Certification, occurs in the field and may require up to 26 weeks. The purpose is to provide facility oriented operational training, leading to Position Qualification and Facility Rating on operational functions and knowledges which are unique to the individual Flight Service Stations.

Some proposed changes in Phase III have been under review. These were occasioned by the FAA program to automate the Flight Service Stations in the near future, and are expected to require some adjustments in the training program. Changes will be based on experience in the existing pass-fail program, together with information about the Flight Service Automated Station (FSAS) equipment and experience with present automated systems. The major changes proposed are elimination of the Broadcast, Flight Data, and Teletype blocks of instruction. A new block, called Systems Data Coordinator, is proposed for supportive skills required to operate within an FSAS environment. Broadcast requirements are to be incorporated in the Inflight and Pilot Briefing blocks. The effects of these changes on the current pass-fail procedure will be evaluated as the changes are implemented.

EVALUATION

The student evaluation procedure is designed to be both an integral part of training and a component of the in-depth screening process. This requires the utilization of measurement instruments that are representative of the job areas found in Flight Service Stations and the inclusion of a sufficient number of measures to permit reliable pass-fail assessment. Further, it must assess both job knowledge and readiness for job performance. Thus, it must include measures of academic classroom achievement and ability to perform job-like tasks in a laboratory environment. Table 1 shows the numbers and types of activities and measurement instruments available for student pass-fail evaluation, by phase of training. Block tests are utilized in Phase 2, and achievement tests, graded laboratory problems, and skills tests, in Phase 3.

The evaluation procedure places primary emphasis (80%) on laboratory problem performance and on assessment of the primary FSS skills of pilot preflight briefing and inflight and emergency services by means of skills tests. These skills are most crucial to the FSS mission and scores on the tests have shown considerable differences between performance levels of FSS controllers and developmental students. The secondary skills of Weather Observer, Teletype, Broadcast, and Flight Data receive a combined weight of 15% and are measured in laboratory exercises. They are supportive of the primary skills, but have shown less discrimination between FSS controllers and developmental students. The academic component of seven block tests (Phase II) receives a combined weight of 5%. Although scores on the Block tests have tended to be high and have not shown predictive discrimination among developmental students, these tests have been included in the weighted score to encourage the developmental students to perform seriously on them.

Phase II

Table 2 shows the means and standard deviations of the Phase II Block tests (expert-prepared multiple choice tests) for a sample of the first 118

190.

Table 1

Measurements available
for
Pass-fail evaluation

Activity	Phase II Block Tests	Phase III Achievement Tests	Phase III Graded Laboratory Problems	Phase III FSS Skills Tests
Broadcast	1			
Teletype	1			
Weather Observer	1*		4	
Flight Data	1		2	
Preflight	1*	3	4	1
Inflight	1		4	1
Emergency Services	1		4	1

*Certification examination by the National Weather Service

191.

Table 2

Means and standard deviations of a sample of FSS students
on Block Tests (Phase II) at the FAA Academy.
N = 118 (Classes 78-01 to 9006).

	Mean	Standard Deviation
Weather Observer	90.87	5.25
Teletype	85.61	8.33
Broadcast	90.10	5.50
Flight Data	86.52	7.27
Preflight	88.37	7.10
Inflight	84.70	6.92
Emergency Services	86.25	7.48

192.

students in the new program. The mean scores on the Block tests were at
the mid-eighty percentage levels and higher and this indicates that the
training objectives for Phase II were being achieved. Attrition from
this phase has generally been rare and for reasons other than proficiency.

Phase III

Phase III, laboratory training, involves an attempt to simulate the
operational environment. It provides the opportunity for each student to
practice what has been learned in the classroom, and is the best available
phase of ATC Academy training for measurement of the ability to perform
the duties of the various FSS operations positions. Ability to perform
those duties is assessed during the accomplishment of laboratory problems
and also by use of FSS skills tests.

Laboratory problems: Four graded laboratory problems are prescribed
for each of the primary Preflight, Inflight, and Emergency Services acti-
vities, as well as four for Weather Observer, and two for Flight Data.
A different instructor scores the student on each problem and without
awareness of other instructors' ratings of the students in other laboratories
or in earlier performance. This procedure involving multiple raters, each
one grading each student independently, adds objectivity to the grading
process. It also provides an excellent defense in cases in which students
may protest failing grades with complaints of bias or unfairness related
to race, religion, or sex.

Problem scores are derived by over-the-shoulder observation, using
checklists developed by task analysis of operational work performance.
The observer indicates whether each step has been successfully achieved,
The checking of behavioral task elements on these checklists tends to
minimize subjectivity in the assessment process. Sample copies of the
forms used in the Inflight position over-the-shoulder evaluations and the
evaluation record and worksheets are presented in the Appendix. Instructor
evaluations have also been used as part of the process of determining student
aggregate scores in the Terminal and EnRoute courses, adding a new element
and increasing the stability of pass-fail assessment on each problem. In-
corporation of the instructor's numerical assessment of each student's per-
formance on a problem and predicted potential performance on future problems,
were also added to the checklists used in the FSS Training Program.

Skills tests. (See also Chapter 12.) Skills tests were developed to
help support evaluations in the laboratory for the critical areas of Pre-
flight, Inflight, and Emergency Services. These tests are scored by objective
keys, which serve to minimize the impact of instructor bias. These new tests
have been administered to samples of developmental students and journeymen at
operational Flight Service Station facilities and of FSS students at the ATC
Academy, for validation, relation to on-the-job performance, and for standardi-
zation purposes. Since norms have been developed to describe performance
of these groups, a new student's scores on the skills tests can be compared
to those of FSS field personnel as well as to those of other FSS students at
the Academy. Thus Pass-Fail evaluations are based on standards derived from

193.

the analysis of actual job performance data. Unacceptable student performances, resulting in failing scores, are determined by reference to the normative data.

The Preflight Briefing Skills Test presents in written form the kind of dialogue that takes place when a pilot communicates by radio or telephone for a briefing. For the Weather Skills Test, the student is supplied with weather data sheets from which to supply this kind of information, and must complete multiple choice questions regarding the appropriate responses to the pilot's questions. The Inflight Skills Test presents in written form the kind of dialogue that takes place between those working this position and pilots who are airborne. The student is provided weather data, a flight service area map, and an action list of 19 possible actions from which to select responses for the questions. As multiple actions should be taken in most situations, the student may erroneously omit some actions that should be identified and include actions that would be inappropriate or wrong. These omission and commission errors seem to be quite independent negative scores or forms of error measurement. The data suggest that a combination of the two scores (omits + wrongs) is an eminently useful score for the Inflight Skills Test. The Emergency Services Skills Test was first a Very High Frequency Omnidirectional Range System (VOR) orientation problem, utilizing a branching technique to present the student with optional paths to follow in locating a lost aircraft. If the student makes a poor decision, opportunities are provided in the form of Minor Error paths, for a return to the better "Major Decision" path. Phraseology questions also were provided in the test. The Phraseology and Major Decisions subscores seemed to be parallel measures of the same skill. A combination of these provides a reliable single measure. The Minor Error path has much logical appeal to specialists in air traffic control and is needed to maintain the simulation. A total score which combined these sub-scores, was recommended for use in determining Pass-Fail for the Emergency Services Skills Test. This test was later expanded into a combination of VOR Orientation, Automatic Direction Finding, (ADF), (Manual) Direction Finding (DF), and Time and Distance problems.

Information concerning the suitability and operational usefulness of the FSS skills tests was obtained empirically by administering them to one sample of 253 practicing Air Traffic Controllers at field operational facilities, who ranged in grade from GS-5 to GS-12 and over (the majority, 169, at GS-11) and in experience from less than 1 year to well over 3 years, and also to a sample of 273 ATC FSS students in 1977-1978. The two samples were fairly similar in demographic and background characteristics, as shown in Table 3.

Detailed data on the relationships between the scores of both groups and other relevant variables were reported by Pickrel (1979). For the field controllers, improvement was found in mean test performance scores, related to: increased experience on the job (e.g., between those with less than 2 years experience and those with over 2 years experience),

194.

Table 3

Comparison of the ATC Academy Sample (N = 273) and the Field FSS Sample (N = 253) on sex, education, and pilot qualification.

	Percentages	
	ATC Sample	Field FSS Sample
	N = 273	N = 253
Sex		
Male	79	87
Female .	20	11
No response	1	2
Education		
MA degree	3	1
BA degree	25	16
College 3-4 yrs	11	19
2 yrs	23	19
1 yr	16	18
HS diploma	22	27
No HS diploma	<1	1
Pilot Certificate		
Instrument	12	7
Commercial	13	18
Private	11	11
Student	5	6
None	59	57
No response		1

195.

advancement in GS grade level (peaking at GS-11), the holding of commercial or instructor pilot certificates, compared to lesser or no pilot experience, training in EnRoute Flight Advisory Service (EFAS), compared with those without that additional training, FAA Academy training (although those who completed Academy training over a year previously exceeded those "old-timers" who never attended the Academy; the "old timers" did exceed those who completed the Academy during the past year), and assignment to a facility having a full-time Evaluation Professional Development Specialist (EPDS), compared to assignment to a facility without such a specialist. No differences were found related to sex.

For the student sample, the FSS skills tests correlated well with scores on the Fundamentals of Air Traffic Control, a multiple choice general information test (.37 to .41), with average laboratory grade (.32 to .38), and with each other (.27 to .42).

The performance of students in the new training program initiated in 1978, on the skills tests, was compared with that of students who studied in the old program. The improved performance in the new program has been noteworthy. For example, the average number of errors on the Inflight Skills Test was lower than in the old program by 17 and total scores on the Emergency Services Skills Test averaged 10 points higher. Cutoff scores were set for each skills test, based on data on 118 students in classes 78-01 to 9006 (the initial classes in the new program). As shown in Table 4 (upper part), these eliminated 10.2% (Preflight), 6.8% (Inflight), and 5.9% (Emergency Services), respectively, of the students in this sample. To facilitate the calculation of phase grades, the raw scores were converted to a scale of 0-100 and the cutoff on each scale was set at a converted score of 70.

PERFORMANCE STANDARDS

Scores of operational personnel and of former students were used to establish performance standards to be required for new students to become eligible for acceptance into the operational facility work force (See Pickrel, 1979). These reflected the general view that the capabilities of a total work force will gradually improve if an entrance eligibility requirement is adopted for new personnel which sets their minimum acceptable test performance at a level that exceeds the performance level of the lowest 5% of a current work force, and this principle was embodied in the empirical procedures followed.

The new standards were adopted as pass-fail criteria in FAA Academy FSS training and were based on detailed analysis of data obtained on 253 active FSS air traffic control specialists at field operational facilities, and 273 students from the last eight classes (78-1 through 78-8) in the "old" training program and the first eight classes (78-01 through 9006) in the "new" training program that was adopted in 1978. In order to complete ATC Academy training, students are now required to meet this standard as evidence of their readiness for an operational assignment.

Table 4

Cutoff scores for skills test and laboratory grades
for ATC Academy students from classes 78-01, 78-02,
and 9001 to 9006 (N = 118)

Skills Test	Type of Scores	Cutoff	Students Below Cutoff	
			Number	Percent
Preflight	Rights	Less than 11	12	10.2
Inflight	Negative: Wrongs plus omissions	Greater than 32	8	6.8
Emergency Services	Rights	Less than 15	7	5.9
Laboratory				
Preflight	Average	Less than 70	8	6.8
Inflight	Average	Less than 72	6	5.1
Emergency Services	Average	Less than 76	7	5.8

197.

The performance of 118 students in the first eight classes 78-01 through 9006, in the new training program were the statistical base for derivation of pass-fail cutoffs.

LABORATORY GRADES

With the development of improved laboratory procedures for the new FSS training program, in 1978, the number of graded problems was increased, instructor evaluations were incorporated into the grading process, and stricter, more objective quantitative grading procedures were introduced to provide an improved scoring base for the identification of weak students. Cutoff scores for the new training program, below which approximately 5% of the students scored, are given in the lower part of Table 4. The cutoffs for passing Preflight, Inflight, and Emergency Services are laboratory problem score averages (4 problems) of 70, 72, and 76, respectively, which would fail 6.8%, 5.1%, and 5.8% of the students in these classes.

FINAL GRADES

The final grade is derived from a weighted composite of all phase grades, and is based on academic block tests, secondary position graded laboratory problems, primary position graded laboratory problems, and primary position skill tests. The sum of these scores, a weighted composite, in then converted to a phase grade. The weighted composites are calculated according to the following linear combination:

5.0% Academic Average (average of all block tests)

15.0% Graded Laboratory problem averages for Weather Observer, Teletype, Broadcast, and Flight Data

15.0% Preflight Laboratory Average (4 graded problems)

15.0% Preflight Skills Test Converted Score (Table 5)

12.5% Inflight Laboratory Average (4 graded problems)

12.5% Inflight Skills Test Converted Score (Table 5)

12.5% Emergency Services Laboratory Average (4 problems)

12.5% Emergency Services Skills Test Converted Score (Table 5)

Each of the foregoing measures is based on a grading metric from 0 to 100.

In order to provide normative data for future classes, all weighted composite scores were computed for classes 78-01 to 9006 according to this linear combination. These were standardized using the mean (82.84) and standard deviation (4.54) of the weighted composite scores. Then they were re-scaled to convert weighted composite scores to final Phase Grade.

198.

FAILURE CRITERIA

A student can fail Phase 3 in two ways, by Position Failure or Phase Grade Failure. A Position Failure occures when a student fails both the skills test and the laboratory problems in one of the primary positions -- Preflight, Inflight, or Emergency Services; any of these is a basis for failure of Phase III. The Phase Grade entry for that student is the word FAIL, with no numerical score. A Phase Grade Failure occures when a student achieves a final Phase Grade below 70; in that case, he or she fails the course. The final Phase Grade is derived from a weighted composite of all phase grades.

SUMMARY OF DATA, ALL VARIABLES

Table 5 presents means and standard deviations (at the right) and intercorrelations of scores on the block tests, graded laboratory problems, and the FSS skills tests for the student population in the new training program. Scores on the graded laboratory problems for the several positions correlated well with each other, and were generally higher than the laboratory problems in the old training program. The measures of student performance in the training program correlated well with the final Phase Grade. This Phase Grade is the best availoble measure for determination of pass-fail. The classroom block test, laboratory, and skills test scores, by position, correlated fairly well with each other and this supports their use as a composite; however, the final Phase Grade is considered to be a superior measure.

The Multiplex Controller Aptitude Test, MCAT, a new measure for use in initial screening of applicants, correlated .49 with the Final Grade. This test was designed to measure aptitude for the type of work being taught in the new training course, and a correlation of this magnitude is a mutually supportive indicator. The final Phase Grade has demonstrated a desirably high relation to student aptitude for this work. Members of the Academy instructional staff who counsel stumbling students during training are allowed access to MCAT scores to judge whether the source of difficulty lies in lack of aptitude for the work. Office of Aviation Medicine and FSS Academy personnel have continued research and development work to refink and strengthen the performance standards.

TABLE 5

INTERCORRELATIONS OF PHASE III TESTS, NEW TRAINING PROGRAM

CLASSES 78-01 to 9006

	Variable		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	N	Mean	SD
	Weather Ober	1	45	35	20	51	30	36	69	56	17	26	25	39	24	25	31	14	-14	-30	-29	10	25	16	25	37	118	90.87	5.25
	Teletype	2		51	24	29	17	28	64	43	39	21	43	53	23	38	29	12	-11	-32	-27	02	22	12	20	41	118	85.61	8.33
	Broadcast	3			19	39	25	24	58	28	16	40	16	29	28	15	28	14	-03	-13	-10	-09	03	19	13	23	118	90.10	5.50
Block	Flight Data	4				25	20	19	52	23	18	36	41	38	18	17	33	24	-15	-06	-15	-02	13	24	21	32	118	86.52	7.27
Tests	Preflight	5					44	52	77	54	-01	37	24	20	30	28	35	35	-25	-17	-30	16	37	28	40	50	118	88.37	7.10
	Inflight	6						24	58	27	05	22	16	18	19	24	25	05	-25	00	-20	14	21	36	34	30	118	84.70	6.92
	Emergency Serv	7							66	32	-11	21	25	17	21	33	42	31	-12	-07	-14	25	39	25	38	46	118	86.25	7.48
	Academic Ave	8								59	19	43	42	51	37	42	50	31	-24	-24	-33	14	38	38	45	60	118	87.48	4.35
	Weather Ob Lab	9									27	37	40	64	27	31	39	27	-36	-29	-46	15	41	26	40	58	118	83.75	8.12
	Teletype Lab	10										22	42	77	21	15	04	-03	-02	-11	-08	-12	11	15	16	21	118	83.44	14.01
	Broadcast Lab	11											38	55	43	39	35	19	-25	-14	-28	01	41	19	37	47	118	82.96	6.22
Laboratory	Flight Data Lab	12												80	29	33	29	15	-16	-10	-19	-05	37	37	45	46	118	76.82	11.92
Grades	Laboratory Ave	13													40	39	33	19	-22	-23	-32	-03	41	34	45	55	118	81.79	7.20
	Preflight Lab	14														52	33	23	00	-05	-03	18	29	-02	17	58	118	80.12	6.66
	Inflight Lab	15															46	31	01	-18	-10	09	30	08	21	62	118	84.36	6.12
	Emer Serv Lab	16																35	-19	-27	-32	18	52	17	40	65	118	87.69	6.30
	Preflight Test	17																	-26	-10	-27	08	23	08	19	59	118	14.64	3.08
	Omissions	18																		01	81	-07	-22	-22	-27	-35	117	14.43	5.07
Skills	Wrongs	19																			60	02	-18	-16	-20	-25	117	6.67	3.72
Tests	Inflight Test	20																				-04	-29	-27	-34	-43	117	21.10	6.31
	Emer Serv Test	21																					13	-07	04	54	118	31.86	9.28
	MEAT	22																						39	84	49	92	76.75	13.54
	OKT	23																							82	16	92	45.64	12.55
	Total(MEAT-OKT)	24																								40	92	122.39	21.82
Final Grade		25																									118	82.32	4.76

FLIGHT SERVICE STATION EVALUATION FORM   INFLIGHT

____ OJT
____ PROGRESS
____ QUALIFICATION   HOURS TRAINING TO DATE_____
____ OVER THE SHOULDER
____ REQUALIFICATION   FOR THE PERIOD OF _____

S-SATISFACTORY   U-UNSATISFACTORY   N-NOT OBSERVED

EVALUATION FACTOR	EXPECTED PERFORMANCE	PERFORMANCE INDICATOR	OBSERVED PERFORMANCE
1. PRIORITY OF DUTIES	ADHERES TO PRIORITY OF DUTIES	1. PERFORMED ALL POSITION FUNCTIONS IN ACCORDANCE WITH LOCALLY PUBLISHED PRIORITY OF DUTIES.	
2. POSITION RELIEF BRIEFING CHECKLIST	PERFORMS POSITION CHECKLIST	1. PERFORMED REQUIRED NAVAID DIAL FUNCTIONS 2. PERFORMED REQUIRED AIRPORT LIGHTING CHECKS. 3. COMPARED CONSOLE INSTRUMENTS AS REQUIRED. 4. MADE TIME CHECK IN ACCORDANCE WITH PRESCRIBED LOCAL PROCEDURES. 5. RESET CLOCKS WHEN REQUIRED. 6. OBTAINED RELIEF BRIEFING.	_____ _____ _____ _____ _____ _____
3. MONITORING FUNCTIONS	ENSURES OPERATING STATUS OF NAVAIDS AND COMMUNICATION EQUIPMENT.	1. PROMPTLY RESPONDED TO AURAL/VISUAL ALARMS. 2. PERFORMED REQUIRED DIALING FUNCTIONS TO RESTORE NAVAID OUTAGES. 3. NOTIFIED MAINTENANCE OF EQUIPMENT MALFUNCTIONS IN ACCORDANCE WITH PRESCRIBED LOCAL PROCEDURES.	_____ _____ _____
4. OPERATE COMMUNICATION AND OTHER POSITION ASSOCIATED EQUIPMENT	1. OPERATE POSITION EQUIPMENT USING PRESCRIBED PROCEDURES	1. USED TRANSMITTERS/ RCVRs SELECTIVELY. 2. USED STANDBY TRANSMITTERS/RECEIVERS IN ACCORDANCE WITH PRESCRIBED PROCEDURES 3. USED LANDLINE COMMUNICATIONS EQUIPMENT IN ACCORDANCE WITH PRESCRIBED PROCEDURES	_____ _____

____ OJT
____ PROGRESS
____ QUALIFICATION   HOURS TRAINING TO DATE_____
____ OVER THE SHOULDER
____ REQUALIFICATION   FOR THE PERIOD OF _____

S-SATISFACTORY   U-UNSATISFACTORY   N-NOT OBSERVED

EVALUATION FACTOR	EXPECTED PERFORMANCE	PERFORMANCE INDICATOR	OBSERVED PERFORMANCE
		4. USED PROPER MICRO-PHONE TECHNIQUE (VOICE LEVEL, MIKE POSITION, HEADSET OPERATION)	
		5. ACTIVATED/DEACTIVATED AIRPORT LIGHTING CONTROLS AS PRESCRIBED.	
	2. OPERATE DIRECTION FINDING EQUIPMENT USING PRESCRIBED PROCEDURES.	1. CHECKED BEARING ACCURACY TO TARGET TRANSMITTER.	
		2. TURNED OFF TARGET TRANSMITTER AFTER BEARING CHECK COMPLETED.	
		3. CHECKED STROBE LINE BEARING SIMULATION SWITCHES.	
		4. RETURNED INNER SWITCH TO OPR POSITION AFTER ALIGNMENT CHECK COMPLETED.	
		5. USED INTERROGATION SWITCH TO STABILIZE BEARING FLUCTUATIONS OF THE STROBE LINE.	
		6. SET DIRECTION FINDER GAIN TO HIGH POSITION TO LENGTHEN A SHORT STROBE LINE RESULTING FROM A WEAK SIGNAL.	
		7. MAINTAINED THE DIRECTION FINDER GAIN IN NORMAL POSITION AT ALL OTHER TIMES.	
		8. USED ALIDADE AS A BEARING REFERENCE.	
		9. MAINTAINED QDM/QDR SWITCH IN QDM POSITION WHEN OBTAINING HOMING INFORMATION.	
		10. USED SQUELCH CONTROL AS PRESCRIBED.	
		11. NOTIFIED MAINTENANCE OF MALFUNCTION IN ACCORDANCE WITH PRE-SCRIBED LOCAL PROCE-DURES.	

_____ OJT
_____ PROGRESS
_____ QUALIFICATION   HOURS TRAINING TO DATE_____
_____ OVER THE SHOULDER
_____ REQUALIFICATION   FOR THE PERIOD OF _____

S-SATISFACTORY   U-UNSATISFACTORY   N-NOT OBSERVED

EVALUATION FACTOR	EXPECTED PERFORMANCE	PERFORMANCE INDICATOR	OBSERVED PERFORMANCE
5. CONTINUALLY ANALYZE METEOROLOGICAL AND AERONAUTICAL INFORMATION	MONITOR AND UPDATE ALL WEATHER AND AERONAUTICAL DATA AS AVAILABLE	1. KEPT WEATHER AND AIRMEN INFORMATION CURRENT.	
6. PIREP SERVICE	SOLICIT, INTERPRET AND DISSEMINATE PIREPs AS PRESCRIBED.	1. SOLICITED PIREPs WHEN CRITERIA EXISTS. 2. DISSEMINATED PIREPs AS PRESCRIBED. 3. PREPARED PIREPs IN PROPER FORMAT.	
7. DETERMINED AIRCRAFT OVERDUE	INITIATES REQUIRED ACTION ON OVERDUE AIRCRAFT	1. INDICATED RECOGNITION OF OVERDUE AIRCRAFT. 2. ATTEMPTED RADIO CONTACT AS PRESCRIBED. 3. INITIATED LOCAL COMMUNICATIONS SEARCH. 4. INITIATED QALQ IN A ACCORDANCE WITH PRESCRIBED LOCAL PROCEDURES.	
8. SERVICES TO EN ROUTE AIRCRAFT	CORRECTLY PROVIDES SERVICES TO EN ROUTE AIRCRAFT IN ACCORDANCE WITH PRESCRIBED PROCEDURES	1. PROMPTLY RESPONDED TO AIRCRAFT CALLS. 2. PROVIDED ALL REQUIRED DATA. 3. USED PRESCRIBED PHRASEOLOGY 4. RECORDED ALL AIRCRAFT CONTACTS AS PRESCRIBED 5. USED PRESCRIBED STRIPMARKING SYMBOLS/ CHARACTERS. 6. PROVIDED HAZARDOUS AREA REPORTING SERVICE AS PRESCRIBED. 7. PROVIDED ADVISORIES/ PILOT BRIEFINGS AS PRESCRIBED. 8. PROVIDED VFR CRUISING LEVEL ADVISORIES AS PRESCRIBED. 9. RELAYED AIR TRAFFIC CONTROL CLEARANCES, ADVISORIES AND REQUESTS VERBATIM.	

203.

FLIGHT SERVICE STATION EVALUATION FORM   INFLIGHT

____ OJT
____ PROGRESS
____ QUALIFICATION   HOURS TRAINING TO DATE_____
____ OVER THE SHOULDER
____ REQUALIFICATION   FOR THE PERIOD OF _____

S-SATISFACTORY   U-UNSATISFACTORY   N-NOT OBSERVED

EVALUATION FACTOR	EXPECTED PERFORMANCE	PERFORMANCE INDICATOR	OBSERVED PERFORMANCE
		10. FORWARDED IFR DEPARTURE, PROGRESS, AND ARRIVAL REPORTS AS PRESCRIBED.	_____
		11. PROCESSED VFR MOVEMENT AND CANCELLATION DATA AS PRESCRIBED.	_____
		12. REQUESTED REQUIRED BACKGROUND INFORMATION AS PRESCRIBED.	_____
		13. ISSUED ALTIMETER SETTINGS AS PRESCRIBED	_____
		14. PROVIDED FLIGHT PLAN SERVICES AS PRESCRIBED.	_____
		15. TRANSFER RADIO CALLS AS PRESCRIBED.	_____
9. AIRPORT ADVISORY SERVICE	PROVIDE COMPLETE AIRPORT ADVISORY SERVICE AS PRESCRIBED.	1. PROVIDED AS PRESCRIBED; WIND DIRECTION/ VELOCITY.	_____
		2. FAVORED/DESIGNATED RUNWAY.	_____
		3. ALTIMETER SETTING.	_____
		4. WEATHER.	_____
		5. OBSERVED/REPORTED TRAFFIC.	_____
		6. CAUTIONARY INFORMATION	_____
		7. DENSITY ALTITUDE ADVISORY.	_____
		8. NOTAM INFORMATION.	_____
		9. AIRPORT BRAKING ACTION.	_____
		10. ISSUED UPON REQUEST; TAXI ROUTES AND TRAFFIC PATTERNS.	_____
		11. INSTRUMENT APPROACH PROGRESS INFORMATION.	_____
		12. TIME INFORMATION.	_____
		13. AIRCRAFT EQUIPMENT INFORMATION.	_____
		14. STATED "NO CONTROL TOWER IN OPERATION" WHEN REQUIRED.	_____
		15. FOLLOWED PRESCRIBED PROCEDURE, CONCERNING REQUESTS FOR SIMULATED INSTRUMENT APPROACHES	_____
		16. USED FREQUENCIES AS PRESCRIBED WHEN PROVIDING ALL ADVISORIES	_____

204.

FLIGHT SERVICE STATION EVALUATION FORM INFLIGHT

___ OJT
___ PROGRESS
___ QUALIFICATION   HOURS TRAINING TO DATE_____
___ OVER THE SHOULDER
___ REQUALIFICATION   FOR THE PERIOD OF _____

S-SATISFACTORY   U-UNSATISFACTORY   N-NOT OBSERVED

EVALUATION FACTOR	EXPECTED PERFORMANCE	PERFORMANCE INDICATOR	OBSERVED PERFORMANCE
		17. PROVIDED SVFR SERVICES IN ACCORDANCE WITH PRESCRIBED LOCAL PROCEDURES.	
10. CIRCULAR SLIDE RULE	CORRECTLY SOLVES PROBLEM USING CIRCULAR SLIDE RULE.	1. CORRECTLY SOLVED PROBLEMS FOR TIME 2. DISTANCE 3. SPEED	___ ___ ___
11. AIRCRAFT ESTIMATES	CORRECTLY DETERMINE AIRCRAFT ETE AND ETA	1. CORRECTLY DETERMINED: ETE ETA	___ ___
12. PROVIDE VOR ASSISTANCE	PROVIDE ASSISTANCE TO AIRCRAFT USING PRE-SCRIBED VOR ORIENTATION PROCEDURES	1. OBTAINED REQUIRED PRELIMINARY INFORMATION 2. ADVISED AIRCRAFT TO REMAIN VFR AT ALL TIMES 3. GAVE WEATHER/ AERONAUTICAL INFORMATION AS PRESCRIBED. 4. VERIFIED AIRCRAFT HDG/ALTITUDE ONLY AFTER HEADING INDICATOR AND ALTIMETER WERE ISSUED. 5. USED PRESCRIBED PROCEDURES TO ORIENT AIRCRAFT 6. PASSED COMPLETE/ ACCURATE DATA TO AIR TRAFFIC CONTROL 7. GAVE ACCURATE HEADING INFO TO AIRCRAFT 8. ACCURATELY FIXED AIRCRAFT POSITION 9. PROVIDED ADDITIONAL GUIDANCE IF REQUESTED. 10. USED PRESCRIBED PHRASEOLOGY	___ ___ ___ ___ ___ ___ ___ ___ ___ ___

_____ OJT
_____ PROGRESS
_____ QUALIFICATION   HOURS TRAINING TO DATE_____
_____ OVER THE SHOULDER
_____ REQUALIFICATION   FOR THE PERIOD OF _____

S-SATISFACTORY   U-UNSATISFACTORY   N-NOT OBSERVED

EVALUATION FACTOR	EXPECTED PERFORMANCE	PERFORMANCE INDICATOR	OBSERVED PERFORMANCE
13. PROVIDE DF ASSISTANCE	PROVIDE DF ASSISTANCE TO AIRCRAFT, USING PRESCRIBED PROCEDURES	1. OBTAINED REQUIRED PRELIMINARY INFORMATION.	
		2. ADVISED AIRCRAFT TO REMAIN VFR AT ALL TIMES.	
		3. GAVE WEATHER/AERONAUTICAL INFORMATION AS PRESCRIBED.	
		4. VERIFIED AIRCRAFT HDG/ALTITUDE AS PRESCRIBED ONLY AFTER HEADING INDICATOR AND ALTIMETER WERE ISSUED	
		5. USED DF EQUIPMENT TO DETERMINE AIRCRAFT BEARING.	
		6. PASSED COMPLETE/ ACCURATE DATA TO ATC	
		7. ACCURATELY PLOTTED DF READINGS/BEARINGS	
		8. GAVE ACCURATE HEADING INFO TO AIRCRAFT	
		9. ACCURATELY FIXED AIRCRAFT POSITION	
		10. INDICATED RECOGNITION OF STATION PASSAGE	
		11. PROVIDED ADDITIONAL DF GUIDANCE AS REQUESTED.	
		12. USED PRESCRIBED PHRASEOLOGY	

NATIONAL AIR TRAFFIC TRAINING PROGRAM
FLIGHT SERVICE OPERATIONS
INFLIGHT LABORATORY EVALUATION RECORD

EVALUATION NUMBER _____

VARIATION NUMBER _____

NAME _____

CLASS _____

DATE _____

INSTRUCTOR ASSESSMENT:

INEFFECTIVE	MARGINALLY EFFECTIVE		EFFECTIVE					HIGHLY EFFECTIVE			EXTREMELY EFFECTIVE	SCORE
	10	11	12	17	18	19	20	21	25	26	27	30	
30%													_____ %

PERFORMANCE RATING:

21% PROCEDURES

Error/Score	$\frac{0}{21}$	$\frac{1}{18}$	$\frac{2}{15}$	$\frac{3}{12}$	$\frac{4}{9}$	$\frac{5}{6}$	$\frac{6}{3}$	$\frac{7}{0}$
Additional Errors:		8	9	10	11	12

_____ %

21% COORDINATION

Error/Score	$\frac{0}{21}$	$\frac{1}{16}$	$\frac{2}{11}$	$\frac{3}{6}$	$\frac{4}{0}$
Additional Errors:	5	6	7	8	9	10

_____ %

14% OTHER

Error/Score	$\frac{0}{14}$	$\frac{1}{12}$	$\frac{2}{10}$	$\frac{3}{8}$	$\frac{4}{6}$	$\frac{5}{4}$	$\frac{6}{2}$	$\frac{7}{0}$
Additional Errors:		8	9	10	11	12

_____ %

7% STRIP MARKING

Error/Score	$\frac{0}{7}$	$\frac{1}{6}$	$\frac{2}{5}$	$\frac{3}{4}$	$\frac{4}{3}$	$\frac{5}{2}$	$\frac{6}{1}$	$\frac{7}{0}$
Additional Errors:		8	9	10	11	12

_____ %

7% PHRASEOLOGY

Error/Score	$\frac{0}{7}$	$\frac{1}{6}$	$\frac{2}{5}$	$\frac{3}{4}$	$\frac{4}{3}$	$\frac{5}{2}$	$\frac{6}{1}$	$\frac{7}{0}$
Additional Errors:		8	9	10	11	12

_____ %

My performance on this problem has been reviewed with me by the instructor.

_____
Developmental Specialist's Signature

TOTAL EVALUATION SCORE _____ %

_____
Instructor's Signature

FAA AC 81-2478   FEB 81

207.

NATIONAL AIR TRAFFIC TRAINING PROGRAM
FLIGHT SERVICE OPERATIONS
INFLIGHT LABORATORY EVALUATION WORKSHEET

NAME _____

CLASS _____

DATE _____

EVALUATION NUMBER _____

VARIATION NUMBER _____

PROCEDURES
P = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

COORDINATION
C = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

OTHER
O = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

STRIP MARKING
S = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

PHRASEOLOGY
PH = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

AIRCRAFT IDENTIFICATION	ERROR CATEGORY	DESCRIPTION OF ERROR

AIRCRAFT IDENTIFICATION	ERROR CATEGORY	DESCRIPTION OF ERROR

208.

AAC-933C

FAA AC-79-XXX

SEP 7

A student who fails BOTH the Laboratory Problems and the Skills Test in the Preflight Position or the Inflight Position or the Emergency Services Positi fails Phase III, Course 50223.

PREFLIGHT POSITION

LABORATORY PROBLEMS:

#1   #2   #3   #4

_____ + _____ + _____ + _____ = _____
(BELOW 280 = FAIL)

SKILLS TEST: = _____
(BELOW 70 = FAIL

INFLIGHT POSITION

LABORATORY PROBLEMS:

#1   #2   #3   #4

_____ + _____ + _____ + _____ = _____
(BELOW 288 = FAIL)

SKILLS TEST: = _____
(BELOW 70 = FAIL)

EMERGENCY SERVICES POSITION

LABORATORY PROBLEMS:

#1   #2   #3   #4

_____ + _____ + _____ + _____ = _____
(BELOW 304 = FAIL)

SKILLS TEST: = _____
(BELOW 70 - FAIL

NAME _____

CLASS _____

STARTING DATE _____

209.

PHASE GRADE WORKSHEET

<u>BLOCK TESTS</u>

BC   FD   TT   IF   PF   ES   WO

___ + ___ + ___ + ___ + ___ + ___ + ___ = ___ ÷ 7 = ___ X .05 = ___

<u>BROADCAST GRADED PROBLEMS</u>

#1   #2   #3   #4

___ + ___ + ___ + ___ = ___   ÷ 4 = ___ X .0375 = ___.

<u>FLIGHT DATA GRADED PROBLEMS</u>

#1   #2

___ + ___ = ___   ÷ 2 = ___ X .0375 = ___.

<u>TELETYPEWRITER GRADED PROBLEMS</u>

#1   #2

___ + ___ = ___   ÷ 2 = ___ X .0375 = ___.

<u>INFLIGHT GRADED PROBLEMS</u>

#1   #2   #3   #4

___ + ___ + ___ + ___ = ___   X .03125 = ___.

<u>INFLIGHT SKILLS TEST</u> = ___   X .125 = ___.

<u>PREFLIGHT (PILOT BRIEFING) GRADED PROBLEMS</u>

#1   #2   #3   #4

___ + ___ + ___ + ___ = ___   X .0375 = ___.

<u>PREFLIGHT (PILOT BRIEFING) SKILLS TEST</u> = ___   X .15 = ___.

<u>EMERGENCY SERVICES GRADED PROBLEMS</u>

#1   #2   #3   #4

___ + ___ + ___ + ___ = ___   X .03125 = ___.

<u>EMERGENCY SERVICES SKILLS TEST</u> = ___   X .125 = ___.

<u>WEATHER OBSERVATION GRADED PROBLEMS</u>

#1   #2   #3   #4

___ + ___ + ___ + ___ = ___   ÷ 4 = ___ X .0375 = ___.

WEIGHTED COMPOSITE SCORE   ___.

NAME_____   CLASS _____

PHASE GRADE SCORE

Chapter 11

CONTROLLER SKILLS TEST

Evan W. Pickrel and Jack M. Greener

The Controller Skills Test (CST) is actually a family of proficien-
cy tests developed originally by personnel from the Office of Aviation
Medicine, the Office of Air Traffic Service, and the Office of Personnel
and Training of the FAA to supplement various other performance and skills
assessments used to evaluate students in the non-radar control laboratory
phase of the new FAA ATC Academy Training Program, implemented in January
1976. The actual weighting of the CSTs in conjuction with other training
measures to assess overall training performance is described in Chapter 9.

The initial CSTs, one each for the Terminal and EnRoute options,
were designed to reflect the application of knowledge and skills devel-
oped during the laboratory phase of training and included three basic
elements for evaluation: (1) application of separation standards in
response to situations described by flight strips and/or controller
charts, (2) responding to or forwarding information that pertained to
coordination with other controllers and (3) other items, such as board
management, timeliness of actions, and phraseology.

Development of the tests involved initially a pool of approximately
100 items on viewgraphs, which were presented ATC personnel at several
Terminals and EnRoute centers in various locations. Analyses of the re-
sulting data led to the construction of an operational, 50-item, multiple
choice, paper and pencil CST for each of the two career options. Subse-
quent enhancements by personnel at the FAA Civil Aeromedical Institute
(CAMI) included expansion of the CSTs to allow for inclusion of an addi-
tional fifty experimental items for the purpose of building a data base
which could be used to develop alternate forms of the tests. Thus, al-
though the current versions of the tests contain 100 items each, only the
original 50 operational items are scored for operational purposes. The
tests have a one hour time limit and utilize a "rights only" scoring
formula.

Two kinds of data were presented as being indicative of the initial
validity of the CST (FAA, Note 1; Dailey & Moore, 1979). The first in-
volved comparison of the CST scores of trainees and controllers with dif-
fering levels of training and experience. Although the actual mean scores
were not presented, one FAA note (1976) concluded:

    Scores on the tests showed substantial differences
    between developmental and full performance level
    controllers in centers and IFR terminals and showed
    substantial differences between groups of develop-
    mental trainees in different phases of ATC training.
    The Controller Skills Tests also have been found to

211.

reflect substantial gains in skills during the Non-
radar Control Lab Course in 1975.

Dailey and Moore (1979) reported similar findings from an examina-
tion of CST performance corresponding to levels of controller experience
for a sample of 226 controllers from six Terminals and four EnRoute cen-
ters:

> On the non-radar subject matter of the prototype
> CST, it took a considerable time for ATC's to
> achieve mastery. Few, even with extensive ATC
> experience, made much of a showing on either
> test until their second or third year on the job.
> By then, however, the facility controllers did
> relatively well on the test. This indicates
> that the CST is relatively independent of prior
> experience and measured skills learned on the
> job. On the written test taken from the Academy
> battery, the peak of performance was reached
> after about 12 to 21 months, and the rate of
> mastery seems to begin to fall off after 22 to
> 39 months' experience on the job. On the CST,
> the peak of plateau was reached by those with
> 22 to 39 months experience on the job, and the
> fall off did not begin until 40 to 69 months'
> experience.

These findings led Dailey and Moore to conclude that the CST does in-
deed measure the application of the subject matter rather than just the
knowledge of it.

A second kind of data used to evaluate the CST consists of a com-
parison of CST scores with concurrent performance on laboratory train-
ing problems. The latter data were also reported in the Dailey and
Moore (1979) paper. Laboratory performance was an average score for
each of four problems, computed from an objective point total for the
problem and a subjective, instructor's rating of problem performance.

Results for a sample of 94 academy trainees in the EnRoute course
indicated that the mean correlation between the CST and four problem
averages was .18, as compared to a mean correlation of .23 among the
four problem averages. The fact that the CST correlated with the prob-
lem averages nearly as well as they correlated with each other was in-
terpreted as indicating that the CST was measuring substantially the
same things as the laboratory problems.

Data for 108 Terminal option trainees indicated that alternate forms
of the CST correlated substantially lower with the laboratory problems
than they corrected with each other (.08 vs. .29). The authors inter-
preted this finding as indicating that the Terminal CST was not measur-
ing the same thing as the laboratory problems; it was believed that the
CST was measuring other aspects of the laboratory phase training which
are least likely to be known or easily learned on the basis of prior ex-
perience.

212.

More recently collected data (Mies, Colmen and Domenech, 1977) have reported that the correlation between the CST and a six-problem lab average was .50 for 454 trainees in the Terminal option and .54 for 473 trainees in the EnRoute option.

The generally positive contribution of the Terminal and EnRoute CST as an additional objective measure of the application of knowledge, procedures, and skills developed in laboratory phases of training, has led to the extension of the skills test concept to other FAA training programs. An experimental controller skills test has been developed for the Radar Training Facility (RTF) program and skills tests in the form of written simulations of laboratory problems have become in integral part of the Flight Service Station (FSS) program. The latter tests are discussed in more detail in Chapters 10 and 12.

The achievements of Tucker (Chapter 12), in extending the skills test concept to actual written simulations of laboratory problems in the FSS program, has in turn led to consideration of applying the same extensions to the EnRoute and Terminal programs. An experimental EnRoute CST was developed that was a paper-pencil simulation of laboratory problems and FAA Academy persons were trained in the development of similar simulations of Terminal problems. These were intended for use in addition to the existing measures, but the completion of this work was postponed at the time of the PATCO strike.

REFERENCE NOTE

1. Federal Aviation Administration. Controller Skills Test.
Unpublished Research Note, 1976.

Chapter 12

DEVELOPMENT OF DYNAMIC PAPER-AND-PENCIL SIMULATIONS FOR MEASUREMENT OF AIR TRAFFIC CONTROLLER PROFICIENCY

Joseph A. Tucker, Jr.

BACKGROUND

The job of Air Traffic Controller is one that involves progres-
sive mastery of successive and more demanding skill levels before full
proficiency is acquired. This process, involving both periods of
training and on-the-job experience, takes several years to complete.
The dropout rate during this developmental period was deemed to be un-
acceptably high by the House Government Affairs Committee, U.S. Congress,
in a 1977 report. As Pickrel (1979) has reported, the FAA response has
been to devise and implement new selection procedures and to adopt per-
formance based screening procedures that can be "used at the FAA Academy
as well as later in the training programs to ensure that unsuccessful
controllers are eliminated early in the training process."

The Pass-Fail evaluation procedures reported by Pickrel are based
on "over-the-shoulder" rating during job simulation exercises performed
in a laboratory and upon "paper-and-pencil tests that simulate the lab-
oratory problem." The purpose of the objective-paper-and-pencil skills
test is to increase the relability of the overall Pass-Fail criterion.

The paper-and-pencil skills tests used for performance evaluation
have had varying test item formats ranging from multiple choice (most
frequent) to selection among alternatives. However, all have required
response to single items or stimulus situations, in contrast to respond-
ing to dynamic stimulus situations, as is required by the job simulation
laboratory problems. In the latter case, the consequences of a previous
response become part of the stimulus for the following response. Since
the dynamic situation is conceptually more valid (job like), the FAA has
investigated the feasibility of using dynamic paper-and-pencil skills
tests as a component of its Pass-Fail Performance Evaluation Program.
This Chapter reports upon the results of that investigation to date.
Sections one through four discuss the concept and theory of paper-and-
pencil simulations. Sections five and six report on and present examples
of the FAA applications that have been tried. Sections seven and eight
present alternative delivery possibilities and alternative uses.

THEORETICAL CONSIDERATIONS

Section One – The Concept

McGuire, Solomon and Bashook (1976) used the term written simula-
tions rather than paper-and-pencil stimulations. These authors found

215.

written simulations to be highly effective when applied to problem solving and decision making tasks. They stated that:

"A written simulation focuses on a total problem. The student assumes an active role as the problem solver and is in full command of the situation, determining both the general approach to follow and the specific activities to engage in. Feedback on the decisions the student makes is neither evaluative nor explanatory. Indeed, it is not even corrective - except in the sense that feedback about changes in the conditions of the problem situation, resulting from the student's decisions, is informative. The branching that occurs is clearly a consequence of another person having imposed additional tasks. In short, written simulation is much more nearly analogous to tightly constructed role playing exercises that rely on training surrogates than it is to programmed instruction of any form. Like role playing, it is dynamic and evolving, as is a real-life experience."

and:

"These characteristics of written simulation are now seen as common to all varieties of simulation, including those that employ the most advanced technology our society can provide. And the benefits of learning and/or being tested in a controlled, but realistic, environment which offers the same challenges as does life itself are now seen as applicable to all areas and levels of education. The wide generalizability and high flexibility of simulation technique commend it to all who teach and test."

The relevance of the written simulation technique to the Air Traffic Controller job can be seen in the following statement about problem solving presented by McGuire:

"In order to cope effectively with any problem, individuals must be able to: gather, process and interpret data; use a variety of resources (including expert advice); order priorities of data seeking and decision making; take appropriate action; manipulate the situation to alter it; monitor the effects of these manipulations, and readjust decisions or actions to respond to changing conditions. While the specific demands of particular situations may modify the relative importance of these several skills, all are potentially involved in competent decision making."

That general statement fits the Air Traffic Controller job exactly. It establishes that there is no conceptual reason to doubt the appropriateness of using a paper-and-pencil simulation for evaluating Air Traffic Controller decision making. The practicability of it is demonstrated in section five.

Section Two - Simulation Theory

Simulation theory applied to human performance postulates that the greater the resemblance of the training or evaluation condition to the

216.

task performance condition, the more effective and valid the training or evaluation will be. This can be restated as a principle, that in theory, the best training is that which is conducted in the task situation or situation that closely approximates it. However, since most training is conducted in situations that differ in varying degrees from real life task situation, it is important to assess the effectiveness of training conducted in less realistic situations in comparison to training conducted under conditions of high realism. This comparison identifies the basic transfer of training issues relevant to simulation theory as identified by Gagne, Foster and Crowley (1948). High simulation requires the trainee to behave realistically in a realistic environment. The two broad dimensions of simulation, then, are stimulus fidelity and response fidelity. Abstraction or divergence from realism can occur along both of these dimensions. The degree of abstraction ranges from completely realistic stimuli and responses to purely symbolic stimuli and responses. The paper-and-pencil simulation often involves a maximum of abstraction along the stimulus dimension while maintaining relatively realistic decision response conditions.

Dimensional analysis, as a feature of the process of operationalizing simulation theory, has been a neglected area of methodological research over the past twenty years. French (1956) investigated problem solving (troubleshooting) training and developed a part-task training device called a Malfunction and Circuitry Trainer. In conjunction with this work, he proposed a number of qualitative dimensions which provide a basis for the classification of training and testing devices that are intended to be used as simulators.

According to French:

"Before generalizations concerning training device characteristics and the corresponding training characteristics can be formulated or tested empirically these device characteristics must be translated from the unique to the general. The first step toward generality is the establishment of the relevant dimensions along which features of each device may be classified.

"Although many additional or alternative categories might be suggested, the following dimensions are offered as being particularly relevant to training characteristics. It is proposed that training devices such as the MAC-1 might be classified in terms of (a) degree of freedom of action, (b) degree of remoteness from the field situation, (c) degree of specificity of operational information, (d) degree of symbolic representation."

These four dimensions include two that are for stimulus fidelity and two for response fidelity. French's work is one of the few systematic attempts at dimensional analysis reported in the simulation literature.

217.

CHART 1

SOME FACTORS RELATING TO SIMULATION

FACTORS	REALIA		Replicas	Mockups	APPROXIMATIONS AND ABSTRACTIONS			
	In Place	In a "Lab"			P.O.V A-V	New P.O.V A-V	Audio-Print	Print

Stimulus
  Central
  Surround
  Backaround								

Response								
Consequences								

218.

CHART 2

Conceptual Plan for a Flight Service Station
Emergency Skills Test

ITEM TYPES

	NO. OF ITEMS	POINT VALUE	TOTAL POINTS
1. RELEVANT - IRRELEVANT INFORMATION Decide whether the questions provided are relevant or irrelevant. These items test procedures.	2	5	10
2. SEQUENCING ITEMS Put the steps provided for each item in the correct order. These items test procedures.	10	1	10
3. ERROR IDENTIFICATION Identify errors in emergency intervention. These items test evaluation skills.	5	2	10
4. MINI-CASES Rank order the following intervention strategies. These items test evaluation skills.	5	4	20
5. WX MINI-CASES What would you do next? These items test performance.	10	2	20
6. BRANCHED ITEMS Simulation. These items test performance.	2	10	20
7. ROLE-PLAYING Simulation Final Assessment.	1	10	10
		TOTAL =	100

Lawrence (Note 1) prepared a taxonomy of simulation alternatives that reflects increasing abstraction from a realistic setting (Chart 1), which accounts for both the stimulus and response dimensions. Long and Tucker (Note 2) developed a conceptual plan for a skills test that includes both conventional and dynamic items (Chart 2). The item types are arranged in an ascending hierarchy moving toward dynamic simulation, which appears at item type six. Items one through five are considered achievement test items but not performance test items since they do not require two or more successive responses with an interspersed altering of the stimulus conditions. All of the item types are of a paper-and-pencil format except for type seven. Both Lawrence, and Long and Tucker suggested that a written simulation represents the maximum abstraction from a realistic setting that meets the minimal performance simulation criteria.

Section Three - Part-Task Evaluation and Training

Most performance situations require the use of a combination of complex human skills, in particular, psychomotor and cognitive skills. "Whole-task" simulation attempts to provide an opportunity for the trainee to practice all task skills realistically. "Part-task" simulation attempts to provide training on critical skills that represent segments of a total task. The assumption is that this will result in "positive transfer" to performance on the whole task. Writing within the context of equipment based tasks, Adams (1957) stated that "In whole-task simulators an explicit attempt is made to provide comprehensive simulation for mission training. In part-task simulators the simulation is explicitly limited to a crucial, difficult portion of the total job." The dynamic simulations of Air Traffic Controller proficiency with which this chapter is concerned are of the part-task type. Emphasis is upon decision making simulation, apart from the use of procedural skills involving communication and psychomotor manipulations.

Section Four - Profiles

All human performance leaves a record or track of an individual's activities. Even in activities that are largely cognitive, such as problem solving, there is a record of the overt steps taken by the performer. Both the process and results can be evaluated against models of excellence (standards).

The evaluation procedure used in laboratory situations is accomplished most frequently by an observer who evaluates performance as it is occurring. The evaluation may also include an overall rating at the end of the performance. Where there is computer monitoring of performance, it is possible, in theory, for the computer to be programmed to provide the evaluation.

Another procedure of particular relevance to paper-and-pencil simulation is to limit the number of records that the performer can produce,

220.

but still allow for dynamic performance. This permits the pre-evaluation of each of the possible records (patterns, profiles) against a model of excellence. This profiling has been used for the ATC paper-and-pencil simulations discussed in Sections five and six. The pre-evaluation process enables the assignment to each of the possible profiles of a value relevant to a criterion. The profiles are unique in that they may not be further subdivided. The total number of unique profiles gives the complete ultimate breakdown of the problem investigated. H. A. Toops (1948) has developed the basic theory of unique patterns.

Tucker (1951) demonstrated empirically that the unique pattern technique can be used effectively for prediction and need not be limited to classification only. He demonstrated that restriction of the number of categories on the score axis - a necessary control for the use of unique patterns - is not a methodological limitation, but rather a desirable empirical practice.

Tucker's research, along with Toops' theoretical postulates, supports the use of unique profiles for personnel evaluation. Since a unique profile can be a representation of a performance under part-task simulation conditions, both research and theory support the validity of using part-task simulation for evaluating purposes.

Section Five - The Flight Service Station Application

The development of the Bass-Fail evaluation procedure for Air Traffic Controllers is discussed by Pickrel (Note 3) in Chapter 10. He selected the Emergency Service job function of the Flight Service Station (FSS) Controller fro the investigation of alternate items and test forms that might be used for skill measurement. Then, Long and Tucker (Note 2) developed seven item types (Chart 2) for review. Item Type 6, a branching performance item, was selected for further development.

Item Type 6 is a profile item (Note 2). The following presents a description and an example of this type of item.

The profile item. In many real-life situations a person must perform to meet two or more conceptually discrete requirements simultaneously. Consider the following excerpt from a communication between the pilot of a small aircraft and an air traffic controller (radio).

Pilot: OKC Radio, Cessna 77466, over.

Radio: Cessna 77466, Oklahoma City Radio, go ahead.

Pilot: Radio, I seem to have gotten stuck on top. Can't find any
        breaks in the clouds to get down. Can you help me? I can't
        fly instruments.

Radio: 466, state your type of aircraft, fuel remaining, and
        weather at your altitude.

221.

Pilot: I'm Cessna 172, have about two and a half hours fuel. Weather is clear above the clouds. Looks like the tops are about 4000, seems almost like a low stratus deck.

Radio: Roger. Remain VFR at all times. Advise me if a heading or altitude change is necessary to remain VFR. PWA altimeter is 2987. While in straight and level flight, reset your heading indicator to agree with your magnetic compass and advise me of your altitude and heading.

Pilot: I'm at 5500 and heading 145.

Radio: Roger. Why type navigational equipment do you have and are you transponder equipped?

Pilot: My VOR doesn't seem to be working right. I have a transponder.

Radio: Squawk 7700 on your transponder. You are north of PWA D/F site. Standby.

Radio: Center, Oklahoma City Radio with a lost aircraft.

Center: Go ahead.

The objective of the air traffic controller was to obtain, quickly and efficiently, information to be give to an Air Traffic Control Center for use in locating the lost aircraft by radar. An evaluation of the air traffic controller's performance could have included:

a) Essential information

   Obtained essential information only,
   no omissions or non-essentials

b) Sequence

   Obtained information in a preferred sequence

c) Phraseology

   Used standard, tested terminology.

The introduction of a preferred or essential sequence as a dimension of evaluation permits a test developer to prepare a performance test from a "real life" case, as the following example using the vignette above demonstrates. The example uses two dimensions -- sequence and essential information. A preferred sequence of information collection by the air traffic controller is assumed to be important.

The following example consists of five pages from a 26-page profile test. The profile test presents all logical variations of the three

222.

CHART 3

Prototype FSS Skills Test
Profiles and Scores

Profiles		Page Choices	Profile Score	Information Score	Total Score
Best Sequence	All Essential Information	1a,2a,5b	9	20c = 1	9 or 10
Average Sequence	All Essential Information	1a,2b,6b 4a,3b,9a	8	20c = 1 27d = 1	8 or 9 8 or 9
Poor Sequence	All Essential Information	1b,3c,10c 1c,4b,12a	7	18d = 1 19d = 1	7 or 8 7 or 8
Average Sequence	1/3 Information Missing	1a,2b,6a 1a,2a,5a 1b,3b,9b	6	14c = 1 8b = 1 13Ic=1	6 or 7 6 or 7 6 or 7
Average Sequence	Unnecessary Information	1a,2b,6c 1a,2a,5c 1b,3b,9c	5	15a = 1 16a = 1 16a = 1	5 or 6 5 or 6 5 or 6
Poor Sequence	1/3 Information Missing	1b,3c,10a 1c,4b,12b 1c,4a,11b,12a 1c,4a,11b,12b 1c,4a,11b,12c	4	21c = 1 13d = 1 19d = 1 13I = 1 15a = 1	4 or 5 4 or 5 4 or 5 4 or 5 4 or 5
Poor Sequence	Unnecessary Information	1b,3c,10b 1c,4b,12c	3	22d = 1 15a = 1	3 or 4 3 or 4
Poor Sequence	Limited Information	1a,2c,7a 1b,3a,8a 1c,4c,13a 1c,4a,11a	2	71c = 1 81c = 1 23b = 1 24b = 1	2 or 3 2 or 3 2 or 3 2 or 3
Poor Sequence	2/3 Information Missing	1a,2c,7b or c 1b,3a,8b or c 1c,4c,13b or c	1	71c = 1 81c = 1 13Ic = 1	1 or 2 1 or 2 1 or 2

223.

statements of essential information in the excerpted dialogue above. A performer can make a minimum of three or a maximum of five page selections along the alternative paths to obtain the essential information to contact the Control Center. Sequence can be evaluated as Best, Average, or Poor. Essential information can be evaluated as Complete, 1/3 missing, 2/3 missing, and/or unessential. These evaluation alternatives can be averaged into nine scoring categories, into one of which each of the 25 performance profiles can be classified. Chart 3 presents the evaluation code for each of the performance alternatives.

The profile 1a, 2a, 5b implies that a performer selected choice a, Page 1; choice a, Page 2; and choice b, Page 5. This "best" profile chosen by the performer replicates exactly the model communication presented in the dialogue above. The sequence arises from the first chosen three pages. Page 20 presents a check item that verifies that the performer knows what information he obtained. Each of the 25 available profiles ends with a variation of the check item. This item is scored separately from the performance item.

The profile 1a, 2c, 7b means that a performer selected only one useful information page but took advantage of the multiple choice format on the check item on Page 7 to report to the Control Center information not obtained by his preceding performance. The value of scores 1-9, ·Chart 3 results from assigning values to each of the possible profile categories.

Application. Subsequently, the item was developed into a prototype skills test - FSS Skills Test, Emergencies 1, Form A. Based on critiques by expert FSS controllers, the test sequence category was dropped and measurement was limited to obtaining information. Chart 4 reflects the changes in structure and scoring based on the critique.

The test was tried out in a pilot study at the ATC Academy using 8 instructors, 17 new graduates, and 31 trainees in the 6th week of training. The results showed a 16% error rate for information and a 23% error rate for the Evaluation item. The test did not discriminate between instructors, graduates, and students. The test did show a range of performance among the subjects. However, it was judged to be too easy for operational use.

Since the pilot study did support the feasibility of using a paper-and-pencil performance test for skills assessment, a decision was made to develop a completely new test. Mr. Jack Nimmo, working with the writer, developed FSS Skills Test, Emergency Services II, a VOR Orientation problem. This test measures decision making and phraseology, and differs considerably in format from Emergency Services I. It includes both fixed and branching sequences. Scoring is done, not by profiles but by assigning a weight of 1 point for each of the 12 phraseology items and 5 points for each of the major decisions items. The test included an option of 3 points for a minor decision, but this was dropped when it proved to be nondiscriminating. It also includes a plotting exercies, but this is not scored.

224.

Example

Situation: Aircraft Lost
                 Pilot not IFR rated or capable

PILOT: OKC Radio, Cessna 77466, over.

RADIO: Cessna 77466, Oklahoma City, go ahead.

PILOT: Radio, I seem to have gotten stuck on top. Can't
                 find any breaks in the clouds to get down. Can
                 you help? I can't fly instruments.

   As RADIO: Which of the following statements would you make
                 next? Place a check in the appropriate blank.

A. _____   466, state your type aircraft, fuel remaining and
                 weather at your altitude.

B. _____   466, remain VFR at all times. Advise me if a
                 heading or altitude change is necessary to remain
                 VFR. PWA altimeter is 2987. While in straight
                 and level flight, reset your heading indicator
                 to agree with your Magnetic Compass and advise me
                 of your altitude and heading.

C. _____   Roger, 466. What type navigational equipment do
                 you have and are you transponder equipped?

PILOT: I'm a Cessna 172, have about 2 1/2 hours fuel.
                    Weather is clear above the clouds. Looks like
                    the tops are about 4000. Seems a lot like a
                    low stratus deck.

RADIO: Check which of the following statements you
                    would make next.

    A. _____   Roger. Remain VFR at all times. Advise me if
                    a heading or altitude change is necessary to re-
                    main VFR. PWA altimeter is 2987. While in
                    straight and level flight, reset your heading
                    indicator to agree with your magnetic compass
                    and advise me of your altitude and heading.

    B. _____   Roger, what type navigation equipment do you have
                    and are you transponder equipped?

    C. _____   Center, Oklahoma City Radio with a lost aircraft.

If you chose a-Go to Page 5.
If you chose b-Go to Page 6.
If you chose c-Go to Page 7.

PILOT: I'm at 5500 and heading 045.

RADIO: Which of the following statements would you make next?

A. _____ Center, Oklahoma City Radio with a lost aircraft.

B. _____ Roger, what type navigational equipment do you have and are you transponder equipped?

C. _____ Roger, state destination and number of people on board.

227.

CENTER: Go ahead.

RADIO: Which of the following statements could you make
          at·this time?

A. _____Cessna 77466, C172, 2 1/2 hours fuel, VFR on top
        lost not instrument rated, request radar help.

B. _____Cessna 77466, C172, 2 1/2 hours fuel, VFR on top
        altitude 5500, heading 045, not instrument rated,
        request radar help.

C. _____Cessna 77466, C172, 2 1/2 hours fuel, VFR on top,
        altitude 5500, heading 045, squawking 7700, D/F
        Bearing 010. Request radar help and D/F help.

PILOT: My VOR doesn't seem to be working right. I
                  have a transponder.

RADIO: Squawk 7700 on your transponder. You are north
                  of PWA D/F site. Standby.

RADIO: Center, Oklahoma City Radio with a lost aircraft.

CENTER: Go ahead.

RADIO: Which one of the following 4 possible statements
                  contains the minimum essential information re-
                  quired by the Center to assist Radio in aiding
                  Cessna 77466.

    a. ( ) Cessna 17466, VFR on top, heading 045, 1 person
          on board, squawking 7700, D/F bearing 010. Re-
          quest radar and D/F help.

    b. ( ) Cessna 77466, C172, 2 1/2 hours fuel, VFR on top,
          altitude 5500, heading 045, squawking 7700, D/F
          bearing 010. Request radar and D/F help.

    c. ( ) Cessna 17466, C172, yellow, 1 person on board.
          2 1/2 hours fuel, VFR on top, altitude 5500,
          heading 045, squawking 7700, D/F bearing 010.

    d. ( ) Cessna 77466, C172, yellow, 2 1/2 hours fuel, VFR
          on top, altitude 5500, heading 045, squawking
          7700, D/F bearing 010, destination Oklahoma City,
          departed Waco. Request radar and D/F help.

CHART 4

FSS SKILLS TEST
Emergencies - 1
Scoring

Selections	INFORMATION	INFORMATION SCORE	CORRECT COORDINATION	MAXIMUM POINTS PER SELECTION
1a, 2a, 5b, 20b	In these sequences all	9	20b 1	9 or 10
1a, 2b, 6b, 19d	information required by the	9	19d 1	9 or 10
1b, 3b, 9a, 17d	ARTCC was collected	9	17d 1	9 or 10
1b, 3c, 10c, 18d		9	18d 1	9 or 10
1c, 4b, 12a, 25d		9	25d 1	9 or 10
1c, 4a, 11b, 12a, 25d		9	25d 1	9 or 10
1a, 2b, 6a, 14a	One third of the information	6	14c 1	6 or 7
1a, 2a, 5a, 8b	required by the ARTCC was	6	8c 1	6 or 7
1b, 3b, 9b, 13b	not collected	6	13c 1	6 or 7
1b, 3c, 10a, 21a		6	21c 1	6 or 7
1c, 4a, 12b, 13d		6	13c 1	6 or 7
1c, 4a, 11b, 12b, 13d		6	13c 1	6 or 7
1c, 4a, 11b, 12c, 15b		6	15d 1	6 or 7
1a, 2b, 6c, 15b	One third of the Information	5	15d 1	5 or 6
1a, 2a, 5c, 16a	required was not collected.	5	16d 1	5 or 6
1b, 3b, 9c, 16a	Information not required by	5	16d 1	5 or 6
1b, 3c, 10b, 22d	ARTCC was collected	5	22b 1	5 or 6
1c, 4b, 12c, 15b		5	15d 1	5 or 6
1a, 2c, 7a	Two thirds of the information	3	7c 1	3 or 4
1a, 3a, 8a	required was not collected.	3	8c 1	3 or 4
1c, 4c, 13a		3	13c 1	3 or 4
1c, 4a, 11a, 23b		3	23c 1	3 or 4
1c, 4a, 11c, 24d	Two thirds of the required information was not collected Information not required by ARTCC was collected	2	24a 1	2 or 3

230.

In May 1977, the test was give a provisional trial at the ATC Academy, using 8 instructors, 7 graduates, and 31 students as subjects. The results showed a mid-point score of 36 for instructors, 28 for graduates and 21 for the students. Scores ranged from 2 to the maximum of 42. The favorable discrimination plus a wide range of scores led to a decision to validate the test for operational use.

The test was administered to 253 FSS controllers at operational facilities and 273 students at the Academy. The results showed a favorable discrimination in favor of the operational personnel and a wide range of performance scores. This permitted the establishment of a reliable lower 5% failure cut-off score for the Emergency Services II Test. Consequently, the test was incorporated into the overall FSS pass-fail evaluation procedure used at the Academy. Pickrel's data (Note 3) for the validation of Pass-Fail performance standards include the data for the Emergency Services II test.

The success of the FSS dynamic paper-and-pencil simulation is empirical evidence of the utility of that type of performance measure as a component of a performance based evaluation system. Next, the feasibility of the technique was investigated for the non-radar EnRoute controller specialty.

Section Six - The EnRoute Applications

The non-radar EnRoute Traffic Control sepcialty is subject also to the pass-fail performance evaluation procedure. The use of dynamic paper-and-pencil simulations might be an important feature of skills testing if the meaningful application of simulation principles were technically feasible. The EnRoute controller handles a complex of activities continuously, rather than handling a single problem to completion, as does the FSS controller. The progress thus far in on-going research on paper-and-pencil simulation of non-radar controller procedures is reported in this section.

This research has focused on three questions: (1) Is it technically feasible to develop practical, meaningful paper-and-pencil simulations for non-radar? (2) If so, to what extent can non-radar procedures be simulated? (3) What scoring procedures are most valid for non-radar skills assessment?

Working with Mr. Robert Van-Gilder and Mr. James Dwyer, ATC controllers, Leesburg, VA. ARTCC, the writer has addressed these issues. After some trail-and-error activity, six paper-and-pencil performance tests were developed, three for the job function of separation, two for coordination, and one for emergency services. The separation and coordination tests begin with a control condition involving four or five aircraft. Each of the six tests requires five control responses from the student. Since there are three response options at each choice point, each test has 243 unique performance profiles - a reasonable, but minimal simulation of the complexity of the non-radar control function. Each of the tests

231.

represents a situation that can be administered as a non-radar proce-
dures laboratory exercise. Consequently, each can be considered a
part-task, mini-case exercise. Since the conceputal difficulty level
of the three tests is low as compared to most laboratory exercises,
stringent scoring criteria are both appropriate and required. Chart
5 presents the first page of a coordination test as an example of the
format.

Using the first separation test, a pilot study was conducted at
Leesburg with a sample of 13 controllers (2 full proficiency, 5 GS-9,
4 GS-7, and 2 developmentals). The scoring procedure was the number
of best choices out of five attempted. The results were that five of
seven full proficiency and grade level 9 controllers had three or bet-
ter best choices; two of six grade level 7 developmental controllers
had three or better best choices, a trend in the direction expected.
Four of the six scoring categories were used. These results encouraged
the authors to continue with the development of the Coordination and
Emergency Services tests. Further data about the tests is discussed
below in conjunction with a review of scoring.

Chart 6 is a copy of the Non-radar Procedures Profile used by in-
structors to evaluate trainee performance on a laboratory exercise.
The Separation paper-and-pencil test covers the evaluation factor,
shown in Chart 6 as Separation-1, Not insured or instructor had to in-
tervene. However, the test has no provision for instructor interven-
tion. For the next category, Coordination and Communication, Item 2,
Coordination is not thorough, is covered by the Coordination test, but
Items 1 and 3 are observational variables that cannot be covered.

For the category, Traffic Management and Control Judgment, Items
1 through 4 are covered in the Emergency Services and Separation test.
Item 5 is partially covered by the present separation test and might
be covered more adequately in a subsequent paper-and-pencil simulation.
Item 6 is not covered. The paper-and-pencil simulation tests are too
limited procedurally for Item 7 to be evaluated effectively.

The next category is Operating Methods and Procedures. In this
category, the Emergency Services test covers Items 2 and 3, and a
future exercise could cover Item 4. While Item 1 can be checked in
principle, discrepancies in flight strip posting procedures among field
locations make it impractical. Validation would be very difficult to
accomplish.

The Category, Equipment, Phraseology and Other, is difficult to
cover. Item 4, Makes Unnecessary Transmissions, is covered by all three
tests. Items 2 and 3 cannot be checked. Items 5, 6, and 7 are not cov-
reed in the present three tests. Item 1, Phraseology, can be measured
in paper-and-pencil simulations, but its coverage complicates greatly
the development of the simulations. It appears preferable to use other
testing procedures to check on recognition of correct phraseology.

232.

CHART 5

Coordination Exercise

AIRCRAFT: Aero Center N4141P over Enid at 1700 level at 80 estimating OKC at 1732 requests to change destination to OKC.

CENTER:

A. N4141P Roger, stand by.

B. N4141P, cleared to Karns via V1, maintain 80.

C. N4141P Roger, cleared to OKC via V1, maintain 80.

CHART 6

NONRADAR PROCEDURES PROFILE

1. Name (Last, First, MI)	2. Date	3. Problem	4. Sector

5. Sector Certification ☐ YES	6. Training Phase	7. Training Hours This Session:	8. Total Hours To Date:

9. Traffic Complexity ("X" one) ☐ Routine Not Difficult ☐ Occasionally Difficult ☐ Mostly Difficult ☐ Very Difficult

EVALUATION FACTORS	PERFORMANCE	REMARKS
SEPARATION 1. Not insured or instructor had to intervene.	☐ ☐	Total
COORDINATION & COMMUNICATION 1. Professional manner is not maintained. 2. Coordination is not thorough. 3. Communication is unclear, not concise.	☐ ☐	Total
TRAFFIC MANAGEMENT & CONTROL JUDGMENT 1. Awareness is not maintained. 2. Poor control judgment is applied. 3. Control actions are incorrectly planned. 4. Positive control of situation is not provided. 5. Prompt action to correct errors is not taken. 6. Aircraft identity is not maintained. 7. Board management is not maintained.	☐ ☐	Total
OPERATING METHODS & PROCEDURES 1. Flight strip postings are incomplete or incorrect. 2. Clearance delivery is incorrect/incomplete/ untimely. 3. Ltrs of Agreement/center directives not adhered to. 4. Handoff procedures incorrectly performed.	☐ ☐	Total
EQUIPMENT, PHRASEOLOGY AND OTHER 1. Standard phraseology is not adhered to. 2. Uses poor voice quality. 3. Speech rate is incorrect. 4. Makes unnecessary transmissions. 5. Equipment status information not maintained. 6. Computer entries are incorrect. 7. Equipment capabilities not fully utilized/ understood.	☐ ☐	Total

Developmental's Signature	Instructor or Evaluator's Signature	Supervisor's Signature
Date	Date	Date

FAA Form 3120 ...

This critique shows that paper-and-pencil stimulations can give very good coverage of the "priority" areas of air traffic control function. For emergency services it can be used to evaluate knowledge of procedures in a realistic context, although it cannot be said to be capable of evaluating emergency services technique over a wide range of emergency situations. Since it is difficult to conceive of meaningful paper-and-pencil simulations of the non-priority areas, it may be necessary to augment the simulation exercises with more traditional test items, in order to form a complete EnRoute skills test.

Scoring of written simulations raises complex issues. One objective is to devise a standardized procedure that can be used from exercise to exercise. This can be done by considering each step in an exercise to be an independent item and scoring it as right-wrong or in the case of three alternatives, by 2, 1, 0 if the alternatives are hierarchical. The five-response format of the three EnRoute simulations lends itself to this type of "achievement" scoring. The item scores can be summed over all three tests. It should be noted that this procedure does not evaluate the sequence or profile.

However, since sequence is important in simulation, it may include performance variance not evaluated if the items are scored as independent measures. One method of scoring for profile is to weight the score on a subsequent item based on the quality of the score on a preceding item. This would weight the profile and still provide a score directly additive across tests.

Since the tests are constructed to have best, acceptable, unacceptable response alternatives at each response point, this weighting procedure is feasible. However, while it weights performance based on the last previous step, it does not weight performance on the total profile. For example, the best alternative at response 4 might be to separate two aircraft, for which a top item score would be given; but, the separation should have occurred not later than response 3.

The difficulty with the scoring procedures mentioned above (which may take a variety of forms) is that they are not sensitive to the subtleties of performance inherent in each exercise. The writer has not yet found a standardized procedure that would fit all three tests acceptably. (The problem would not be recognized by developing a standardized procedure using one test only). What was needed was a scoring procedure that would evaluate performance in the same way that an "over-the-shoulder" laboratory instructor would do it, one that would be responsive to the range of performance quality inherent in the test. As a result, the writer set aside the search for a "directly additive" scoring procedure and investigated the possibility of obtaining performance scores for each test individually and then developing a procedure for summating scores across tests.

An analysis of each paper-and-pencil test problem by James Dwyer resulted in a listing of the key activities of each and an identification

of the sequence points by which each activity should have occurred; unfortunately, the criteria differed widely from test to test. Next, a two-stage scoring procedure was devised for each test. Stage 1 set a minimum Pass performance standard. Then, in Stage 2, each of the 243 profiles was given a score based on the standard. The result was a weighted scoring procedure for each test that clearly separated pass from fail, and allowed superior and very poor performance to be identified. Since each test is considered to be of equal performance, a profile of performance over a series of tests (i.e., six or more) could be developed and a percentage passed cut-off criterion set.

To acquire data for checking the performance characteristics of the three tests, two classes (36 subjects) of newly graduated (Phase III) EnRoute controllers were administered the tests as the ATC Academy in July 1979. They, along with two developmental controllers tested at Leesburg, provided test results upon which technical decisions could be made. One issue was: How many simulation tests can be included to form an overall skills test that could be administered during a 1-hour testing period? The Separation test had a median of 6 minutes, with a range from 4-10 minutes. The Coordination test had a median of 3 minutes, with a range of 2-8 minutes. The Emergency Services test had a median of 3 minutes, with a range of 2-7 minutes. Actually 6 minutes was sufficient time for the last two tests with few exceptions. These results suggested that up to eight tests could be administered in a one hour period.

The tests were administered by two instructors without difficulty. The instructions took not over 5 minutes to present, including distribution of materials. The trainees had no difficulty with the materials. During debriefing a large majority indicated that they found the tests challenging, fair, and more acceptable than conventional tests. The results suggested a need for some revisions in the structure of the tests, particularly in the Coordination test, and these have been made.

The question of the degree of similarity between achievement scoring and performance scoring has been checked using all 55 subjects who have taken the separation test under contolled administrative conditions. The achievement score was based on a 3-2-1 allocation to each response for a maximum of 15. The performance score was based on 5 points for a correct decision with a maximu score of 25. The Pedrson product moment correlation between the two was .71, which indicates about 50% common variance. The common variance should be close to 90% for the two scores to be equivalent.

An analysis completed for the Coordination test also indicated differences between achievement and performance scoring. The performance scoring for the Coordination test provides five possible scores 0, 5, 10, 15, and 20, with 15 considered passing (1 error). An analysis of 9 profiles that provide a score of 20 showed them to be functionally equivalent since each required 4 correct decisions out of 4 possible. A 2, 1, 0 achievement scoring procedure was checked for functional equivalence, with the result that some equal achievement scores (a score of 8, for

example) were assigned to profiles that were not qualitatively equal, using performance scoring (some with 1 error, others with 2). This meant that tracks receiving a failing score of 10 using performance scoring would pass using achievement scoring. This supports the writer's belief that achievement type scoring is not appropriate for these exercises.

Following the investigations of the utility of paper-and-pencil simulation exercises for EnRoute and Terminal applications, an effort was made to determine whether FAA Academy course developers could devise such exercises cost effectively. In July 1980, the writer conducted a two week workshop at the FAA Academy with course developers as participants. The workshop demonstrated that course developers readily acquired the "know-how" to develop the simulations. Improvements were made in the prototype non-radar exercises and new ones were planned. In addition, procedures were developed for the preparation of radar simulations for both EnRoute and Terminal applications. This increased greatly the generality of the technique for air traffic controller evaluation and training.

By the summer of 1981 plans were complete for developing both radar and non-radar skills tests and training simulations. In July 1981, the writer conducted a training workshop for five development teams each consisting of a course developer and an education specialist. Two instructors from the Predevelopmental Phase II program at the University of Oklahoma attended also. These teams were working on the development of paper-and-pencil simulation exercises when the occurrence of the Professional Air Traffic Controllers (PATCO) strike resulted in personnel reassignments that terminated the development work for the present.

Another application planned for paper-and-pencil simulations was to use them to provide Pre-developmental controller trainees with simulated job experiences to aid them in making a choice of assignments to the Terminal, EnRoute, or Flight Service career options. However, following the PATCO strike, the entire Pre-developmental program has been closed down for up to one year. Currently, research and development continues, to develop these simulation procedures and to investigate applications using a computer based instruction mode. The more general term, micro-simulation is being used to identify this work, which is moving beyond the paper-and-pencil application.

Section 7 - Delivery Alternatives

Up to this time, paper-and-pencil formats have been used with the FSS tests and the prototype EnRoute tests. A limitation of this procedure is that a student may be tempted to "peek ahead" to check a response. This breaks the "simulation set" that the students should be following. Although this behavior is controlled in part by scrambling the responses and by tight time limits, the potential exists for student non-conformance.

237.

McGuire recommended the use of a latent-image response sheet that requires a student to respond before he can get information as to how to proceed. However, the coding of 243 responses would be a formidable task for the EnRoute simulations.

Other available modalities include oral, audiovisual and computer assisted. (Chart 1) Any paper-and-pencil simulation can be arranged for presentation by a variety of information retrieval devices including the auto-tutor and the new Vis-a-com devices. Also, they can be presented by any computer assisted instruction system. These alternatives are feasible if already in place for other reasons. Generally, the cost-effective decision is to use an inexpensive latent-image response sheet to insure response simulation.

Section 8 - Conclusions

It is a curious historical fact that the initial use of part-task trainers and written simulations was for evaluation rather than for training and other uses. Both French and McGuire referred to this fact from quite disparate settings. Crowder's work (1954) leading to the development of branching programming was preceded by his research on the development of performance troubleshooting tests. The training potential of the testing instruments has been recognized by all developers. Often a training application has been developed, based on a test, because of student demand. The EnRoute controller graduates included in the summer 1979 testing stated that they would have like to have had the tests as training exercises followed by a critique. It is possible that the ATC Academy may include such simulation exercises among its curricular materials.

The development of written simulations must be based on a thorough task analysis of controllers' procedures. The anaysis directed by the writer provides considerable insight into the conceptual activities of the performing controller. The analytical procedures can be used in basic research into controller functions.

In the future, it is possible that part-task simulations can be developed that can be a feature of the initial assessment process. Such simulations could be developed using the stimulus approaches included in the MCAT tests.

One can conclude that written simulation can be an important technique in the study, initial assessment, evaluation and training of Air Traffic Controllers.

238.

REFERENCE NOTES

1. Lawrence, K. A taxonomy of simulation options. Fairfax, VA, 1976.

2. Long, L., & Tucker, J. A. A conceptual schema for a skills test.
Washington, D. C.: Catholic University of America, School of Education, 1978.

3. Pickrel, E. W. The new FSS training program: Performance standards for pass-fail determination. Interoffice Memorandum, Office
of Aviation Medicine, Federal Aviation Administration, Washington,
D. C., 1978.

Chapter 13

POST-TRAINING CRITERION MEASURES IN VALIDATION
OF CONTROLLER SELECTION PROCEDURES

Jack M. Greener

INTRODUCTION

The preceding four chapters have addressed criterion measures based on performance during and up to the end of Academy training. In this chapter attention is drawn to criterion measures based on performance on the job at various career points subsequent to completion of Academy training. Ideally validation of a selection program should include both training level and post-training criteria, the latter including job performance at the most highly qualified level of the job. In the studies that have been carried out, attention has been given to both the relationships among different criterion measures at each level (training and post-training) and also to the relationships between training and post-training measures. The latter are of particular importance because significant relationships between training and post-training performance were assumed in the final validation of the new test battery adopted by the FAA in the fall of 1981 and described in Part IV of this book.

Measures of controller performance used as criteria for prediction are integral to the validation of selection instruments because the basic design of validation requires the establishment of predictor-criterion relationships. Performance measurement also serves other purposes such as the basis for administrative decisions regarding job assignments and promotion decisions and feedback to employees for employee development purposes. Criterion measures are commonly distinguished on the basis of whether they reflect performance during training or post-training. In addition, both types of measures may also be subjective or objective.

During-training criteria usually focus on the rate and extent to which students learn job-related information, while post-training measures emphasize competence in typical performance of the job. In some cases, the latter has been approached in terms of global indications of the value of an employee to the organization with respect to one or more dimensions of performance. In general, during-training criteria have been better predicted by preemployment tests than post-training criteria.

Although performance during training may on occasion be regarded as an end in itself, as when reduction of high training costs is an issue, the use of during-training criteria for validation of selection instruments usually implies the assumption that the training criteria are correlated with post-training criteria at an acceptable level.

The distinction between subjective and objective criteria refers primarily to the relative objectivity of the performance recording and scoring process. Objective measures lend themselves to clear cut and ready determination of response classification and scoring, while subjective measures require observers to apply personal interpretation and judgment in scoring or evaluating individual responses. Examples of subjective criteria include observer and supervisor ratings while examples of objective criteria include structured response tests, counts of individual production output and measures of employee attrition. Objective criteria are usually favored over subjective criteria since they are subject to less bias and measurement error.

A review of the major FAA studies involving post-training controller performance measurement since 1956 indicates that a considerable range of possible performance measures has received consideration over the years. Major types of measures indicated include:

    attrition
    disciplinary action vs no disciplinary action
    medical complaints vs no medical complaints
    amount of sick leave taken
    job progression - time to promotion recommendation
    job progression - from low complexity to high complexity facilities
    peer ratings of controller proficiency
    supervisor ratings - experimental ratings
    supervisor ratings - administrative ratings
    observer ratings - "over-the-shoulder"
    computer-derived measures - simulator based

The number of studies reviewed was fairly large, but they tended to be based upon a smaller number of data sets collected at various times. The central criterion issues of reliability and criterion interrelationships received extensive coverage in only one or two studies related to each data set. It is noteworthy that the most frequently used measures were attrition and experimental (not administrative) ratings by supervisors.

DESCRIPTION OF MEASURES

Attrition. Attrition has been the most widely used criterion measure in ATCS selection research, extending from early research by Trites (1961) to the present, as reported by Boone in Chapter 9. The popularity of attrition as a criterion measure appears to reflect three factors. First, it is a rather visible expense item in the cost of maintaining the public air traffic control system and cost estimates for recruiting and training of replacement controllers, in a system with a prevailing attrition rate of 25% to 40%, have led to recommendations that the FAA develop special programs to reduce attrition costs (Pickrel, 1979; Mies, Colmen, & Domenech, 1977). Second, much of the FAA research reflects an underlying belief that attrition results mostly from poor performance due to inadequate ability to perform critical traffic control tasks and

242.

that improved aptitude screening would reduce attrition. Third, measures of attrition are more readily and accurately obtained than other measures of job performance.

Although it might appear that attrition is a very straightforward and objective criterion measure, its use in FAA selection research has not been without definitional problems, particularly since most of this research has tended to focus on the prediction of attrition due to poor performance. For example, most studies of attrition during initial Academy training have tended to exclude individuals who withdrew for personal reasons unrelated to poor performance. In addition, the Academy attrition rate has obviously been influenced by the manner and degree to which Academy pass-fail standards have changed over time.

By contrast, measures of post-Academy attrition have rarely been adjusted for separation from the active controller ranks for reasons such as resignations unrelated to poor performance or promotions from active controller status to higher level management positions. The reasons for not attempting to do so appear to be two-fold: first, post-Academy performance has tended to be assessed in a much less precise manner than Academy performance, with the result that it has been more difficult to determine whether attrition in any particular case was due to poor performance; and second, the time and expense involved in conducting the review of individual performance records would have reduced substantially the feasibility of even conducting such a study.

At least one study (Trites, 1961) has implicitly addressed the issue of whether attrition should be defined as "no longer with FAA" or more narrowly as "no longer working as an active controller," by actually including both measures in the same study. Although the actual correlation between the two measures was not reported, both had similar correlations with supervisor ratings obtained four years earlier. One major study of ATCS performance measurement (Trites, Miller & Cobb, 1965) used "separation from FAA" as the attrition measure, but more recent studies have tended to define attrition as "no longer employed by FAA in an ATCS job code" (Cobb, Mathews, & Nelson, 1972; Mathews, Collins, & Cobb, 1974; Mies, Colmen, & Domenech, 1977).

The study by Mathews, Collins, and Cobb (1974) focused on the reasons for attrition and included data from job-exit forms, telephone interviews, and questionnaires. Only the reasons related to family matters appeared to be relatively stable from job-exit to later follow-ups.

Disciplinary action, sick leave, and medical complaints. One early study by Trites (1961) included disciplinary action, sick leave, and medical complaints as criterion measures. These data were supplied by the chiefs of the facilities where the 187 controllers in the study were assigned. Disciplinary action was scored dichotomously to represent the occurrence or nonoccurrence of disciplinary actions taken against

243.

the controller as a result of violations of air traffic rules or procedures. Medical complaints were scored dichotomously as "No Symptoms vs Symptoms" of medical complaints of controllers which were known to and reported by their facility chiefs. Sick leave was scored as the mean number of hours of sick leave taken during the years 1957 through 1960.

Progression. Although not widely used as a criterion measure in ATCS research, job progression has been included in at least two studies (Brokaw, 1959; Mies, Colmen, & Domenech, 1977). The early study by Brokaw (1959) defined job progression in terms of time to recommendation for promotion and included recommendation to first promotion and recommendation to second promotion as separate criterion measures. His data analyses revealed that there was little variation among individuals on either measure and this led him to conclude that the nearly standard timing of these recommendations on the basis of longevity (or seniority) limited their value as criterion measures.

Mies, Colmen, and Domenech (1977) utilized a progression criterion based on job complexity rather than rate of progression. They postulated four general levels of operational complexity among the ATC facilities, ranging from Flight Service Station (FSS) at the least complex level, to Visual Flight Rules (VFR) Terminal facilities at the next complex level, to Instrument Flight Rules (IFR) Terminal facilities, to EnRoute Centers at the most complex level. Progression was scaled dichotomously, with a "1" assigned to individuals who, at the time of the study, were working in a less complex facility than their initial assignment and a "2" assigned to those who were working in a facility of the same or higher complexity level than their initial assignment. Since both initial assignment and subsequent transfers have always been highly influenced by available vacancies, as well as personal preference, the progression measure could not be considered to be a strong measure of job proficiency. Nevertheless, the rationale underlying the use of the progression criterion was that the controllers who progressed to or remained at more complex facilities provided a better return on the FAA's selection and training investment.

Peer evaluations. Peer evaluations were collected as criterion measures in two FAA studies which focused on the relationships between age, job experience, and job performance of controllers. The first was a major project which included a variety of objective and subjective measures of controller performance and actually resulted in several separate publications utilizing the same samples and criterion measures (Cobb, 1967; Karson, 1967; Buckley, O'Connor, & Beebe, 1969).

Cobb (1967) used an anonymous coworker's nomination procedure, in which all controllers, from assistant controller to crew chief at each of four major EnRoute Centers, were requested to nominate the four individuals whom they considered to be "most outstandingly effective" and four whom they considered to be "next most effective." A score of 2

244.

was assigned for each "most effective" nomination and a score of 1 assigned for each "next most effecitve nomination." The nominations were standardized within each facility and correlated about .55 with supervisor ratings for a sample of 525 controllers.

The second study (Mathews & Cobb, 1974) involved a sample of 613 journeyman controllers from 17 radar-equipped Terminals and included experimental ratings on a 7-point scale by supervisors, crew chiefs, and co-workers, who were in each case familiar with the work on the individual to be rated. Analysis of correlations among the ratings within each group indicated that the highest interrater correlations occurred among those by immediate supervisors (.61), the next highest among those by crew chiefs (.46), and the lowest among those by peers (.39).

Supervisor ratings. Supervisor ratings have been used widely as criterion measures in FAA controller selection research. There are several reasons for the popularity of supervisor ratings: (1) Supervisors' ratings have considerable face validity, nothwithstanding their obvious subjectivity; (2) They are relatively easy and inexpensive to collect; and (3) There has been a general lack of suitable alternatives.

Although the FAA adopted a formal field performance review system in the mid-1960's, which included periodic "over-the-shoulder" proficiency evaluations as well as more global administrative evaluations, recognized deficiencies in the administrative evaluations for use as criterion measures have generally led to the collection of special (not for official use) ratings for research purposes. Several studies (Cobb, 1967; Milne & Colmen, 1972; Henry, Kamrass, Orlansky, Rowan, String, & Reichenback, 1975) have noted deficiencies in "over-the-shoulder" ratings, including the fact that those that have been used were not scored quantitatively and were designed primarily for diagnostic and remedial purposes. Such forms which were originally designed for other purposes were generally not suitable for criterion purposes.

One significant trend in the use of supervisors' ratings as criterion measures has been in the direction of increased behavioral specificity of the factors rated. The contrast between the rating forms used by Trites, Miller, and Cobb (1965) and Mies, Colmen, and Domenech (1977) is quite striking. Trites et al. included 14 items on a 5-point scale, plus 2 items answered yes or no, while Mies, Colmen, and Domenech used 54 items on a 7-point scale, plus 4 additional items in varying formats, pertaining to overall evaluation. The form used in the latter study contained nine major categories of rating factors, such as knowledge, judgment, traffic management techniques, and performance under stress, with several items in each category, while the items in the Trites et al. form generally corresponded to the category level of the Mies et al. form. An illustration of this contrast is shown in Table 1.

The trend toward behavioral specificity of rating factors was primarily the result of an effort to improve the objectivity and reliability of the ratings. This effort was based on the belief that if

245.

Table 1

Comparison of two supervisor rating forms illustrating two degrees of specificity

Low Specificity (Trites et al., 1965)	High Specificity (Mies et al., 1977)

Low Specificity (Trites et al., 1965)

1. Steady attention to work and conduct.
2. Ability to organize and make most effective use of time, equipment, and information currently available.
3. Demonstrated attitude and character.
4. Rate of continued improvement.
5. Ability to understand and apply controller procedures.
6. Ability to make decisions required by his position.
7. Display of good judgment.
8. Emotional stability under pressure.
9. Demonstrated aptitude for air traffic control activities.
10. Potential for continued emotional stability in air traffic control activities.
11. Ability to get along well with others.
12. Ability to work cooperatively with others.
13. Present performance of OJT duties (complete only for trainees).
14. Potential ability to perform journeyman duties (complete only for trainees).
15. Do the controller activities of this individual ever have an undesirable effect on air traffic safety? Yes-No
16. If you were a facility chief, would you want this individual on your staff as an active controller? Yes-No

High Specificity (Mies et al., 1977)

KNOWLEDGE
1. Of Procedures.
2. Of Equipment.
3. Of Rules, letters of agreement, etc.

PERCEPTION
4. Detecting critical aircraft, radar targets, obstacles.
5. Identifying aircraft and radar targets.
6. Interpreting display data.
7. Scanning efficiently -- towers, boards, radar.
8. Attending to operational status of equipment.
9. Observing weather and updating weather reports.
10. Detecting critical or potentially hazardous local weather conditions and taking appropriate action.

COMPREHENSION
11. Getting, holding and regaining the picture.
12. Scanning flight strips for significant data.
13. Making required computations of speed, direction and distance.
14. Keeping up with air traffic data.
15. Making accurate assumptions about air speed, altitude, position.
16. Comprehending significance of a situation; seeing a problem developing; distinguishing between routine and potentially dangerous situations.
17. Storing current, short-term air configuration and flight plan data with minimal needs for memory aids.
18. Retaining long-term information on rules, regulations, procedures.

COMMUNICATIONS
19. Command of standard ATC phraseology.
20. Communicating concisely outside of standard phraseology.
21. Clarity of radio and telephone voice and delivery.
22. Understanding and interpreting radio messages.
23. Expressing self orally.
24. Giving instructions or information to pilots or others.
25. Acknowledging radio or verbal communication.
26. Using teletypewriter or other equipment for maintaining contact.
27. Detecting and reporting hazardous local air traffic conditions to the proper control sector.

JUDGMENT
28. Planning which balances all relevant factors for operational decisions.
29. Coordinating in a timely and effective manner.
30. Initiative in taking unilateral action as required.
31. Anticipating future status of air space affecting decisions.

TRAFFIC MANAGEMENT TECHNIQUES
32. Applying ATC procedures.
33. Balancing factors of safety, orderliness, expeditiousness.
34. Command of situations.
35. Operating radar, beacon and/or other equipment.
36. Entering flight data into flight strips or computer as required.
37. Receiving, processing and delivering flight plan information.
38. Servicing flight strip printer.
39. Coordinating FDP problem with error referral position and controller, as required.
40. Reviewing and correcting flight plan as needed.
41. Disseminating relays or delivering operational information to appropriate receivers.

PERFORMANCE UNDER STRESS
42. Handling unusual situations such as emergencies.
43. Functioning under periods of short-term heavy traffic.
44. Functioning under periods of protracted heavy traffic.
45. Speed of reacting in reaching decisions.
46. Ability to function effectively in periods of alternating light and heavy traffic conditions.

INTERPERSONAL SKILLS
47. Assuming role as a member of a team.
48. Accepting instructions or suggestions.
49. Manner of making suggestions to others.

OTHER SKILLS
50. Being aware of, and compensating for, own limitation in performance of operational duties.
51. Maximizing own strengths in performance of operational duties.
52. Making efforts toward self-improvement in performance of operational duties.
53. Displaying responsibility toward job and ATC service in performance of operational duties.
54. Adapting to changes in operational policies, procedures, etc.

OVERALL
55. How willing would you be to assign this Specialist to an operational position where traffic was extremely heavy?
56. How often does this Specialist have an undesirable effect on air traffic operations?
57. Would you be willing to promote this employee to a higher level Specialist position or to a higher density facility?
58. Taking into account all the factors specified in the above questions, how would this Specialist rank relative to all Specialists of like grade and level you have known?

246.

the dimensions to be rated were presented in terms of more specific observable behaviors, raters would know more precisely what they were supposed to observe and evaluate and would respond accordingly with more precise evaluations.

Simulation and Computer-derived measures. Simulator studies of air traffic control problems have been reported in the literature since the 1950's. However, most of the early research was directed toward the evaluation of effects of work-load variables and changes in control procedures on overall system performance and therefore was not directly relevant to individual performance assessment. Boone, Van Buskirk, and Steen (1980) provided an excellent review of this early air traffic control simulation research.

Buckley, O'Connor, and Beebe (1969) reported what in many respects must be considered a benchmark study in applying dynamic simulation of air traffic control problems in the evaluation of controller performance. This study included both a full-scale dynamic simulation of the National Aviation Facilities Experimental Center (NAFEC) Model A simulator and a mini-simulation called the Controller Decision Evaluation (CODE), which did not require use of the Model A simulator.

In addition to the objective, simulator-based performance measures, various other measures were collected, including "over-the-shoulder" observer evaluations and physiological stress indicators. The special purpose supervisor and peer ratings described in the Cobb (1967) study, discussed earlier, and recent official Employee Appraisal Record (EAR) ratings were also included for comparison purposes.

The CODE procedure was not developed to be a representative simulation of the traffic controller's job, but rather was seen as an encapsulation of the most basic component critical to successful controller performance, the ability to detect potential air traffic conflictions. Buckley et al. (1969, p. 6-1) described the CODE procedure as follows:

> The basic concept of the technique involves running a film of a radar display portraying air traffic. The subject controller is asked to pretend he is observing another controller and the other's manner of handling traffic. His task is to indicate when action should be taken to avoid a confliction. He of course must scan the changing traffic pattern to detect possible and developing conflictions. Failure to indicate that action is required reflects his failure to detect the confliction. In this manner, the technique concentrates on the perceptual and decision-making element of the controller's job. Since the traffic flow is preprogrammed on the film, the correctness of his decisions can be objectively determined and scored.

247.

Initially the sample for the CODE analyses was to include all 36 controllers from the full dynamic simulation study; however, equipment problems reduced the CODE sample to 18 of the original 36 journeyman EnRoute controllers. Two basic criterion measures were collected: number of conflictions correctly identified and number of non-conflictions incorrectly identified as conflictions. Although the sample was small and the run-to-run reliabilities were rather low for the two lower traffic densities, reliabilities in the .50-.60 range were obtained for the highest traffic density and the CODE scores correlated .50 and higher with both field ratings and scores on the full-scale simulation criteria.

The results of this study were considered promising both for performance evaluation and controller selection applications.

The full-scale dynamic simulation, in contrast to the CODE simulation, actually allowed the controller to direct the activities of a sample of simulated air traffic, performing characteristic functions, such as ordering changes in aircraft speed or flight path and limiting the number of aircraft in the sample sector at particular times by rejecting traffic handoffs from adjacent sectors. Data were collected on a substantial number of potential simulator indices of controller performance, including 12 single measures and 16 ratio measures based on combinations of the single measures, as shown in Table 2. In addition, there were 12 composites computed from combinations of the single and ratio measures.

Utilizing factor analytic procedures, Buckley et al. concluded that the fundamental dimensions of system functioning could be measured adequately with the following nine measures:

Conflicts per aircraft handled
Conflicts per delay
Delays per aircraft handled
Delay time per aircraft scheduled
Aircraft time in system per aircraft scheduled
Proportion of complete flights scheduled actually completed
Contacts per aircraft handled
Communication time per contact
Proportion of aircraft scheduled that were handled.

Buckley and his associates at NAFEC, aided by improved simulation technology (involving conversion from analog to digital simulators) have continued to refine the simulation procedures in a series of small sample studies (Buckley, 1976; Buckley and Rood, 1977). In 1978, Buckley, House, and Rood outlined the research design for a large scale field study to develop a normative data base necessary for the validation of simulator performance parameters that could be operationalized for training and field performance evaluation. Unfortunately this study was not funded.

248.

Table 2

Dynamic Simulation Performance Measures
(Buckley, O'Connor, & Beebe, 1969)

___

Single Measures

Number of Conflictions
Number of Delays
Delay Time
Number of Aircraft Delayed
Aircraft Time in System
Aircraft Time in System for Completed Flights
Flight Time Deviation for Completed Flights
Number of Completed Flights
Number of Control Instructions
Number of Contacts
Communication Time
Number of Aircraft Handled

Ratio Measures

Number of Conflictions/Number of Aircraft Handled
Number of Conflictions/Number of Delays
Number of Delays/Number of Aircraft in Sample
Delay Time/Number of Delays
Delay Time/Number of Aircraft Delayed
Delay Time/Number of Aircraft in Sample
Number of Aircraft Delayed/Number of Aircraft in Sample
Aircraft Time in System/Number of Aircraft in Sample
Aircraft Time in System for Completed Flights/Number of Completed
    Flights
Flight Time Deviation for Completed Flights/Number of Completed
    Flights
Number of Completed Flights/Number of Completed Flights Scheduled
Number Control Instructions/Number of Aircraft Handled
Number of Contacts/Number of Aircraft Handled
Communication Time/Number of Aircraft Handled
Communication Time/Number of Contacts
Number of Aircraft Handled/Number of Aircraft in Sample

___

RELIABILITY OF POST-TRAINING CRITERION MEASURES

Concern about the reliability of the ratings provided by observers, instructors, and supervisors has been expressed repeatedly over the course of controller research from 1950 to the present. Henry et al. (1975) provided an extensive review of research of the reliability of ratings of controller performance up to the time of their study, and only one reliability study has been reported since their review. Apparently, this has been a result of a new FAA operational procedure, in which supervisors rotated shifts with the same crew and thus qualified alternate evaluators were not available to provide the multiple evaluations of the same person required for interrater reliability analysis (Milne & Colmen, 1972).

Estimates of the reliability of ratings have frequently been difficult to compare across studies because some reports have failed to specify clearly whether coefficients reported were average correlations (Pearson r) or correlations corrected for the number of raters by application of the Spearman-Brown prophecy formula. As noted by Ebel (1951), if the issue at hand is the reliability of individual ratings, one whould use the average correlation among the ratings of several raters, whereas if the concern is the reliability of a composite rating that combines ratings from several individual raters, the average correlation among the individual ratings should be adjusted for the number of raters to determine the reliability of the composite. One should also note that Spearman-Brown corrected reliability estimates are not comparable across studies unless they are based on the same number of raters per ratee.

Another issue in the comparison of reliabilities from study to study is whether a reliability estimate is an index of interobserver agreement regarding performance observed at the same time by all raters or whether the reliability estimate is based on ratings of performance observed at different times. If multiple observers provided independent ratings of performance for the same work sample, the reliability coefficient would reflect the reliability of the observation and evaluation process among the raters. However, if the ratings were obtained from observers who observed individuals performing at different times, then the reliability coefficient would be influenced both by unreliability due to observer disagreement and unreliability due to variations in individual performance from one observation period to the next; in that case, the two components of unreliability would be confounded and it would not be possible to estimate their relative influence.

Nagay (1950) conducted a study in which 48 senior controllers observed and evaluated 42 controllers in nonradar control for several weeks. He reported reliabilities of .43 when supervisors observed the same controllers at the same time and .22 when the supervisors observed the same controllers at different times. The ratings were based on observed instances of outstanding and unsatisfactory controller behavior on 23 elements of specific, predefined control behaviors.

Brokaw (1959) reported data indicating an average intercorrelation of .69 among the independent summary ratings by four FAA instructors of the performance of a sample of 20 ATC students. The raters in this study were given training in rating theory and used a 5-point scale with nine items, covering proficiency, ability, attitude, and emotional stability.

In a follow-up of the Brokaw study, Trites (1961) collected ratings by mail from supervisors on a subset of 149 controllers, using a form similar to that used in the Brokaw study. He reported a Spearman-Brown corrected reliability of .75 based on the ratings of those controllers who received ratings from two or more different supervisors. The proportion of the total sample of 149 controllers who were included in the reliability analyses was not reported.

Trites, Miller, and Cobb (1965), in a study of the criterion measures used in FAA selection research during the early 1960's, reported reliabilities for supervisor ratings made at about the same time versus reliabilities of ratings made up to several years apart. The uncorrected correlations between two average ratings made 10 to 12 months after academy graduation were .58 for 468 EnRoute controllers and .78 for 262 Terminal controllers, while the uncorrected correlations between the average in the first year ratings and average ratings one or two years later were .43 for 367 EnRoute controllers and .47 for 244 Terminal controllers. The ratings were based on the sum of the first 14 items given in the Table 1, rated on a 5-point scale.

Cobb (1967) reported results from a study that included a comparison of ratings by the same supervisors, using items of a more general nature, designated by Form B, versus ratings using items of a more specific and technical nature, designated Form C. Although the actual items used were not reported, the 14 items in the general form (Form B) were apparently very similar to those reported in the first part of Table 1, while the 33 items in the specific form (Form C) were the same as those in the semi-annual "over-the-shoulder" field rating from in use at the time, but scaled to provide a quantitative indication of performance. Both forms used a 5-point scale to indicate level of performance. Based on samples of approximately 300 controllers with multiple ratings, Cobb reported average intercorrelations of .35 for Form B and .29 for Form C. He considered the observed reliabilities to be lower than expected and commented that some raters were not as well informed as others (Cobb, 1967, p. 5):

> In those instances where an individual was the recipient
> of multiple ratings, only one represented an evaluation
> by his immediate supervisor; others were by other super-
> visory personnel who may have been less knoledgeable re-
> garding the individual's proficiency. In retrospect, it
> became obvious that the procedure should have included
> the collection of such ratings from only the immediate
> supervisors.

251.

Another study by Mathews and Cobb (1974) provided additional data in support of Cobb's (1967) observation of the importance of using raters who are most familiar with work of each controller rated. This study compared the reliabilities of ratings by supervisors with the reliabilities of ratings by crew chiefs, who were one level of supervision removed from the controllers. With a sample of over 100 Terminal controllers, reliabilities of .62 were obtained for ratings by supervisors, compared to .46 for ratings by crew chiefs.

None of the reliability studies reviewed to this point have actually focused on the situation where two or more observers provide an evaluation of an individual observed in a very focused sample of controller behaviors, which is the explicit condition necessary for a meaningful "over-the-shoulder" rating. Data appropriate for evaluation of the reliability of over-the-shouulder ratings are scarce, but three studies have reported some relevant data.

Henry et al. (1975) summarized the "new" controller performance rating procedures developed by the System Development Corporation (SDS) under contract to the FAA during the period from 1970 to 1974. In addition to reviewing significant features of the SDC rating sytem, which were designed to contribute to improved quality ratings, they reported interrater reliabilities of .65 and .67 for over-the-shoulder evaluations in two field studies; this information was provided in a private communication from one of the SDC researchers.

The Buckley, O'Connor, and Beebe (1969) study described in the previous section investigated the performance of 36 journeyman controllers, using a dynamic computer simulation at NAFEC. In addition to obtaining objective, computer-based measures of performance (reported later in this chapter), each controller was observed and rated by over-the-shoulder procedures by three observers on two one-hour simulation runs at each of three simulated traffic densities. For these ratings, Buckley et al. reported interobserver agreement coefficients (intraclass correlation) for four of nine items rated plus traffic confliction errors as determined by observer judgment (Table 3). It is interesting to note (1) that the overall evaluation was considerably more reliable than the more specific technique-oriented items, (2) that interobserver agreement was relatively unaffected by variations in traffic density, and (3) that interobserver agreement with respect to overall performance was in the same range as their agreement in judging conflictions. The interobserver agreement reliabilities for the overall evaluation were .59, .76, and .67 for three increasing traffic densities. When these agreement coefficients are adjusted for the number of observers (3), the corresponding reliability estimates are .81, .90, and .86 for the respective densities.

In contrast to the data just reviewed, which address the issue of the reliability of the observation and evaluation process for multiple observers observing the same individual at the same time, additional

Table 3

Interobserver Agreement Coefficients
for Dynamic Simulation Observer Ratings
(Buckley, O'Connor, & Beebe, 1969)

| | Intraclass Correlation Coefficients | | | | | |
| | Density 2 | | Density 3 | | Density 4 | |
Measure	Run 1	Run 2	Run 1	Run 2	Run 1	Run 2
Technique Involves Risk Taking	.43	.40	.28	.38	.30	.20
Technique Involves Excessive Caution	.11	.40	.36	.30	.33	.33
Separation is Assured	.45	.46	.64	.57	.58	.52
Overall Evaluation	.60	.57	.75	.77	.70	.64
Number of Conflictions	.63	.72	.70	.77	.81	.70

data were collected to address the issue of the reliabilities of observer ratings and system measures on the same subjects from one simulation trial to another. In this case the reliability estimates for the ratings include the influence of both interobserver agreement and variations in individual performance from one time to the next while the reliabilities of the system measures, with the exception of observer-judged conflictions, include only the influence of variations in individual performance over time.

Table 4 contains illustrative reliability data for selected observer rating and system performance measures. Reliabilities varied considerably for various rating dimensions and the overall evaluation was more reliable than most of the more specifically focused rating items. Some of the rating items, such as "technique involves risk-taking," had relatively low reliabilities which declined considerably with increasing density, while others such as "prompt performance" were fairly reliable and generally unaffected by variations in traffic density.

The reliabilities of the simulator system measures tended to be more variable than those of the rating measures with respect to both differences between measures and differences from one traffic density to another for the same measure. Since reliabilities of the system measures tended to increase with increases in traffic density for some measures and to decrease with increases in traffic density for other measures, combining the system measures into composites tended to attenuate the effects of the two opposing trends. Thus the reliabilities of the system measure composites were more stable across the varying levels of traffic density.

Buckley et al. reported median corrected reliabilities of .63, .65, and .78 for the 12 single simulation measures listed in Table 2, compared to median corrected reliabilities of .75, .70, and .56 for the 9 over-the-shoulder observer ratings. In general, the reliabilities of the objective simulator measures increased with increases in traffic density, while the reliabilities of observer ratings declined. However, it should be noted that the highest traffic density was judged to be more difficult than that typically encountered under operational conditions.

The most significant findings of the Buckley, O'Connor, and Beebe study were: (1) Controller performance varied considerably from one trial to the next, (2) Objective simulator performance measures were not much more reliable than the subjective over-the-shoulder observer ratings, and (3) Two independent simulator trials of one hour duration each were required to produce reliabilities in the .70 range for both objective simulator measures and subjective observer ratings.

Boone, Van Buskirk, and Steen (1980) reported a simulation study comparing over-the-shoulder ratings with computer-derived measures, in an effort to evaluate the feasibility of using computer-derived measures

Table 4

Comparison of Run to Run Reliabilities for Selected
Observer Ratings and Simulator System Measures
(Buckley, O'Connor, & Beebe, 1969)

Observer Ratings

	Density 2		Density 3		Density 4	
	r	r_c*	r	r_c	r	r_c
Technique involves Risk-taking	.53	.69	.24	.39	.11	.20
Separation is Assured	.49	.66	.54	.70	.52	.68
Prompt Performance	.64	.78	.59	.74	.58	.73
Overall Evaluation	.69	.82	.51	.68	.35	.52

System Measures

	r	r_c	r	r_c	r	r_c
N of Conflictions	.60	.75	.62	.77	.34	.51
N of Delays	.26	.41	.43	.60	.67	.80
Flt Time Deviation for Completed Flts	.68	.81	.48	.65	.46	.63
N of Aircraft Handled	.06	.11	.36	.53	.72	.84
2-Variable Composite	.54	.70	.62	.77	.38	.55
12-Variable Composite	.66	.80	.59	.74	.59	.74

Note: the 2-variable composite includes No. Conflicts / No. Aircraft Handled and Delay Time / No. Aircraft in Sample; the 12-variable composite includes all 12 single measures listed in Table 2.

*r_c indicates the Spearman-Brown corrected reliability of the performance measures averaged over two one-hour simulation runs.

to evaluate student performance in simulated air traffic control problems. Twenty-four EnRoute and 24 Terminal students tested at the NAFEC Dynamic Simulation Facility on five radar problems of increasing complexity received instructor evaluations following the standard format described in Chapter 9, for evaluation of laboratory problem performance. These included over-the-shoulder ratings of a problem average score (tabulation of specific errors) and instructor assessment of performance for each problem, a total score combining the problem average and instructor assessment for each problem, and a global rating to the student's potential for becoming a full performance level radar controller. Reliability coefficients for the over-the-shoulder evlauations and global ratings are presented in Table 5. The reliability coefficients reported for the problem average were actually agreement indices, based on the ratio of the number of agreements to the total number of errors recorded by two instructors during simultaneous observation of the same student, rather than conventional correlations of the total number of errors recorded by each instructor for the same student. In addition, the problem average reliabilities appear to have been based only on observations of problems 4 and 5, while the reliabilities for the instructor assessments and global ratings were intraclass correlations across all five problems. Thus, these data did not provide a direct comparison of interobserver reliabilities and situational stability reliabilities on the same measures.

RELATIONSHIPS AMONG CRITERION MEASURES

One central issue to be considered when different criterion measures are used to assess performance, either in the same study or in comparing one study to another, is the relationships among the various measures. In some cases different measures are thought to be measures of essentially the same underlying performance dimensions or comparable alternate measures of overall performance and thus are expected to be highly correlated. In other cases, various measures are thought to be measuring different and possibly independent dimensions of performance and thus are not expected to be highly correlated.

A related issue is the stability of individual job performance across career stages; if relative performance were known to change considerably from one career stage to another, then it would be very important to specify the career stage for which selection instruments are to be validated.

Trites (1961) reported correlations among training grades, instructor ratings, and supervisor ratings of job performance, within the first year after training. As can be seen in Table 6, all of the correlations were significant, positive, and in the low to moderate range; correlations were generally higher between similar measures collected closer to the same point in time. Table 6 also shows a correlation of .28 between supervisor ratings collected in 1957 and the absence of disciplinary action during the period 1957 through 1960.

Table 5

**Reliability Coefficients for the Over-the-Shoulder
Evaluation and Global Rating by Option
(Boone, Van Buskirk & Steen, 1980)**

	Problem Average	Instructor Assessment	Total Score	Global Rating
Terminal	.33	.58	.43	.23
EnRoute	.29	.56	.43	.27

Table 6

Correlations Among 1957 and 1961 Criterion Measures
(Trites, 1961)

	1957 Lecture Grade	1957 Instructor Rating	1957 Supervisor Rating
1957			
Lecture Grade	-		
Instructor Rating	.49	-	
Supervisor Rating	.33	.59	-
Promotion Recommendation	*	*	.18
1961			
Supervisor Rating	.24	.45	.33
Active vs. Inactive Controller	.26	.24	.24
With FAA vs. Not With FAA	*	.16	.20
No Disciplinary Action vs. Disciplinary Action	*	*	.28

Note - All entries are significant at the .05 level.   N=145

* Data not available

Table 7 presents correlations between training performance measures and later job performance measures, as reported by Trites, Miller, and Cobb (1965). Academic grades, laboratory grades, and instructor ratings all had generally modest but significant correlations with supervisor ratings obtained approximately one year after Academy graduation (supervisor rating-1) and supervisor ratings obtained up to three or four years following Academy graduation (supervisor rating-2). Post Academy attrition (separated from FAA) had a low correlation with instructor reservations about a trainee's ability to become a controller, for both EnRoute and Terminal samples, and was also correlated with laboratory performance for the EnRoute sample. In addition, the correlations between supervisor ratings obtained several years apart were .43 and .47 for the EnRoute and Terminal samples, respectively.

Buckley, O'Connor, and Beebe (1969), in the simulation study reported earlier, had access to recent official supervisor ratings (EAR) and the experimental field ratings (global - Form B, specific - Form C, peer nominations) collected by Cobb (1967), as well as the observer ratings and objective simulator measures collected on their sample of 36 controllers. The correlations among the summary scores for these measures are shown in Table 8. Although the overall pattern of these correlations is somewhat erratic and limited by both the small sample and the limited reliabilities of some of the measures, there appears to be a common thread of moderate positive correlations among the various measures. The field ratings on Form C, with the more specific item format, generally had the lowest correlations with the other measures. Cobb (1967) reported correlations in the .50 range between the peer nominations and Form B and Form C ratings for his larger sample, compared to correlations in the .30 range for the same variables in the Buckley, O'Connor, and Beebe sample, possibly indicating a slight restriction of range sample bias in the latter study.

A study by Mathews and Cobb (1974), which included ratings of several hundred Terminal controllers by peers, supervisors and crew chiefs, indicated a range of correlations among the three sources of ratings, from .56 to .59.

Chiles and West (1974), in a study of the validity of the CAMI Multiple Task Performance Battery described in Chapter 4, reported a correlation of .41 between FAA Academy instructors' estimates of each trainee's potential to become a fully rated air traffic control specialist and retention as a controller 2 to 2½ years later for a sample of 214 trainees. A correlation of .29 was obtained for post-Academy retention even when the 10 Academy failures were excluded from the sample.

Mies, Colmen, and Domeneck (1977) in a major validation study discussed in Chapter 18, Part 11, used an aggregate success criterior which included measures of training performance, supervisor ratings, progression, and attrition. The only correlation reported among the individual measures was .10 between progression and supervisor rating. However, the part-whole correlations of the four respective individual measures with the aggregate were .93, .94, .36, and .72, respectively.

Table 7

Correlations Between Training Performance and
Later Job Performance
(Trites, Miller, & Cobb, 1965)

	Academic Grades	Laboratory Grades	Acad. & Lab. Grades	Instructor Rating	Instructor Reservations
Supervisor Rating-1					
EnRoute	.19**	.29**	.28**	.33**	.15*
Terminal	.31**	.07	.23*	.24**	.11
Supervisor Rating-2					
EnRoute	.08	.20**	.17**	.34**	.16*
Terminal	.30**	.19**	.29**	.27**	.22*
Separated from FAA					
EnRoute	.06	.16**	.14**	.02	.11*
Terminal	.02	.00	.01	.04	.14*

N ranges from approximately 125 to 675.
*p<05
**p<01

260.

Table 8

Correlations Among Objective Simulator Scores, Simulator
Observer Ratings, and Field Performance Ratings
(Buckley, O'Connor, & Beebe, 1969)

	EAR	Peer Nomination	Rating Form B	Rating Form C	Simulator Observer Rating
Simulator Observer Rating					
Density 2	.13	.29	.39*	.11	-
Density 3	.25	.30	.39*	.24	-
Density 4	.26	.22	.19	.00	-
Objective Simulator Composite					
Density 2	.30	.26	.12	-.03	.54*
Density 3	.03	.35*	.20	.08	.52*
Density 4	.01	.28	.33*	.25	.29
Peer Nomination	.54*	-	-	-	-
Rating-Form B	.34*	.35*	-	-	-
Rating-Form C	.31	.30	.87*	-	-

N=36

* p<.05

261.

Boone (Table 5, Chapter 9) compared relationships between Academy laboratory performance and field performance ratings 1½ to 2 years following Academy graduation for EnRoute trainees who entered training between May 1977 and January 1978. The relationship presented in that table corresponds to a correlation in the low .40 range.

SUMMARY AND EVALUATION

Although numerous post-training criterion measures have received consideration in air traffic controller selection research since the 1950's, attrition and supervisor ratings have been predominant in use. Objective computer-derived measures have received experimental attention in simulation studies and have shown promise in feasibility studies but, at present, have not been sufficiently developed and validated for operational use.

Considerable effort has been put forth to refine supervisor rating procedures, presumably to increase their objectivity and reliability. These efforts have led to more emphasis on rating scales with large numbers of very specifically focused items to be completed (ideally) under conditions where the supervisor directly observes the controller "over-the-shoulder" while he or she engages in traffic control activities.

Reliability studies have not yielded particularly impressive results for either the over-the-shoulder format or the more traditional global rating format. However, one should recognize that many of the studies were conducted under conditions that tend to produce low reliability estimates. In particular, the use of alternate raters who are not familiar with the work of the controller and the comparison of ratings obtained several years apart are likely to reduce reliability indices. In more recent studies, reliability analyses have not even been conducted as a result of FAA operational procedures which have supervisors and subordinates on the same shift rotation schedule and thus preclude the availability of well qualified alternate raters.

Studies of criterion interrelationships have shown low to moderate correlations between training performance and later job performance, moderate correlations among job performance ratings from peers, supervisors, and crew chiefs, and moderate correlations between supervisor ratings obtained several years apart. Although post-Academy attrition has shown only low correlations with either training performance or supervisor ratings, it has been shown that post-Academy attrition is substantially reduced by increasing Academy attrition through the use of more stringent Academy pass-fail standards.

The implication of the findings reviewed in this chapter are discussed in the following chapter.

Chapter 14

OVERVIEW OF CRITERION MEASUREMENT IN
CONTROLLER SELECTION RESEARCH

Jack M. Greener

INTRODUCTION

The preceding chapters in Part III have described the research and
development and practices in the measurement and evaluation of controller
performance during and after training and have reviewed the FAA research
in which such measures were used as criteria in selection studies. This
chapter considers the nature of criterion measures used in selection
validation research and presents an assessment of those utilized in
studies by and for the FAA. It also provides a critical view of new
measures that have become popular, at least in discussion if not in prac-
tice, as a result of progress in computer-electronic technology.

ISSUES IN CRITERION DEVELOPMENT

Criterion Measures and Performance Assessment

It is widely recognized that the evaluation of employee performance
serves many purposes within organizations and is an essential requirement
in some organizational functions. Among the most commonly recognized
purposes of performance evaluation are: (1) to serve as a basis for ad-
ministrative decisions, such as employee compensation and promotion,
(2) to provide feedback to employees for employee development, (3) to
serve as criteria in studies to evaluate personnel selection and train-
ing programs, and (4) to serve as criteria in programs to evaluate work
methods and procedures.

The extent to which any specific performance measure could serve
adequately more than one of the purposes outlined must be evaluated in
the context of particular circumstances. In general, however, perform-
ance measures designed and collected for specific purposes are usually
not equally appropriate for other purposes. The appropriateness of a
performance measure for any one purpose may even vary depending on the
specific objective or question at hand. For example, the "number of
traffic delays" may be a useful measure in the evaluation of a training
course in air traffic control procedures, but this measure might not
provide any indication of the degree to which trainees actually used the
techniques taught in the course. It is important that the investigator
determine explicitly whether the objective is to examine the impact of
the training course on controller management of traffic delays or whether
the objective is to determine the extent to which trainees learn to ap-
ply the techniques taught. Use of the traffic delays measure as a cri-
terion to answer the latter question would be appropriate only if the
techniques taught in the training program were the sole methods avail-
able to reduce traffic delays.

263.

Many of the considerations in evaluating performance measures for
criterion purposes also apply to performance measures used for other pur-
poses, and in view of the cost of collecting good performance evaluation
data, it is desirable to have systems of data collection which serve more
than one purpose. However, whenever performance assessments are used for
administrative purposes, supervisors may be under much pressure to weaken
the evaluation process because of the potential consequences to the in-
dividuals evaluated.

Characteristics of Criterion Measures

Within the field of personnel psychology the following five considera-
tions have been advocated as a basis for evaluation of criterion measures:
(1) reliability, (2) freedom from contamination or bias, (3) relevance,
(4) freedom from deficiency, and (5) practicality.

Reliability. In its simplest form, reliability refers to the con-
sistency or repeatability of a measure, or in other words, whether in-
dividuals tend to obtain the same scores on different occasions. If
scores or ratings were to vary markedly from one measurement occasion
to another, the measure would be of limited use for determination of the
degree to which individuals actually exhibit the characteristic or be-
havior being measured. For most personnel applications the absolute
amount of variability from one measurement occasion to another is not of
major concern, provided that individuals maintain the same relative stand-
ing with respect to each other over repeated measurements. Thus the cru-
cial consideration is whether the amount of variability between indivi-
duals on spearate occasions exceeds the amount of variability within in-
dividuals across occasions.

Unreliability in measurement may be caused by numerous factors, in-
cluding deficiencies in the measurement device itself (such as problems
of calibration or lack of precision), deficiencies in the measurement
process (such as lack of standardization of the measurement procedure),
and actual changes in the behavior of the individuals measured. Changes
in the behavior of individuals may be transitory and short term, such
as variations due to motivation, anxiety, distraction, and fatigue, or
they may be relatively enduring, long term changes in the circumstances
of employees' lives, jobs, or capabilities (e.g., the acquisition of new
knowledge or skills).

Freedom from contamination or bias. The issue here is the presence
of systematic bias. The basic question to be asked is: "To what extent
does the performance measure under consideration reflect influences
other than the efforts and ability of the individuals whose performance
is being measured?" Examples of contaminated criterion measures include
production rates in situations where individuals do not have comparable
equipment, sales volume for non-equivalent customer bases, accident
frequency under conditions of differential exposure or risk, and per-
formance ratings which are differentially influenced by level of job
responsibility.

264.

In essence, a criterion measure would be considered totally free from contamination only if each individual being evaluated had an equal opportunity to achieve the maximum score. System output measures from which individual performance is inferred appear to be much more susceptible to contamination than measures that focus on actual individual job behaviors.

Relevance and freedom from deficiency. These considerations generally go hand in hand. In assessing the relevance of a proposed criterion measure one wishes to determine that the proposed measure assesses one or more significant, non-trivial aspects of performance germane to the objective or purpose for which it is to be used. Deficiency addresses the completeness of the criterion with respect to the objectives being assessed and requires one to ask whether there are other relevant and significant aspects of performance which should be included.

Practicality. Smith (1976), among others, has suggested the additional requirement that criterion measures be practical in terms of availability, plausibility, and acceptability to the individuals in the organization who will use them in decision making. Certainly it is easy to recognize that some excellent measures of job performance, in terms of the previous four criteria, might require very costly data collection procedures and inordinate amounts of employee and supervisor in terms of the potential consequences of the decisions being made rather than mere administrative convenience.

Dimensionality

One of the most perplexing problems in criterion development and performance evaluation is that of the dimensionality of job performance. The term "job performance" is commonly used as though it represents a unitary concept. Indeed, many of the objectives of performance measurement imply a need to distinguish between varying levels of overall job performance or to distinguish between acceptable and unacceptable job performance. Nevertheless, it is widely recognized in the field of personnel psychology that performance of most jobs actually consists of a number of intertwined dimensions of performance that are not easily represented as a single variable (Ghiselli, 1956; James, 1973; Ronan and Prien, 1966; Smith, 1976).

Thorndike (1949) introduced the theoretical concept of the "ultimate criterion," representing the complete final goal of a selection program, to be differentiated from intermediate criteria and immediate criteria in terms of scope of job activities covered and the point in time in an employee's job tenure that the measure becomes available. The ultimate criterion was seen as being a complex and multidimensional construct that would rarely be directly measurable. Actual validation criteria would by necessity nearly always be substitute, partial, intermediate or immediate criteria related to the ultimate criterion on a rational

265.

basis. Immediate criteria were seen as partial criteria representing the first significant early indicators of an individual's training or job success to provide a basis for evaluation of the validity of a selection program in order to enable needed selection decisions about new applicants during the sometimes lengthy time period before other intermediate partial criteria become available at later stages during training or on the job.

Ghiselli (1956) further elaborated the complexity and multidimensionality of criteria by explicating three separate aspects of the dimensionality of criteria. Static dimensionality was proposed as the aspect which describes the various kinds of possible measures (e.g., speed of work, errors, absences) which could be used to measure performance on a particular job. Dynamic dimensionality was the term used to describe the phenomena where relative competence on a particular performance measure changes over time. Individual dimensionality, the third aspect of criterion dimensionality proposed by Ghiselli, recognized that the contributions made by different individuals could be of equal value to the organization and yet be made along qualitatively different dimensions. In the latter case, it would be inappropriate to use the same dimensions to evaluate everyone's performance.

Ronan and Prien (1966) in a review of the empirical literature of investigations of the interrelationships among various performance criteria for the same job, reported results which document the contentions of Thorndike (1949) and Ghiselli (1956) that, in fact, different criterion measures for the same job are frequently not highly intercorrelated and that correlations between the same criterion measure collected at different stages of employee careers frequently decline over time. The obvious implications of these findings for selection research are that (1) different criterion measures are not readily interchangeable or substitutable for one another, (2) predictors which show validity in relation to one criterion measure should not necessarily be expected to have comparable validity for another criterion measure, and (3) predictors which have validity for predicting job performance at one career stage should not be assumed to be valid predictors of job performance at some later stage. The need for empirical demonstration at different career stages is therefore clear.

Criterion Validity

The considerations of criterion relevance, deficiency, and multidimensionality have led a number of researchers (Guion, 1976; James, 1973; Wallace, 1956) to argue convincingly that meaningful achievements in the validation of selection programs require examination of the validity of the criteria as well. Validation of criteria usually requires a construct validation approach, in which criterion measures are treated as theoretical constructs which are represented by sets of relationships among various dimensions of job performance as well as relationships between various predictors and various measures of job performance.

266.

Guion (1976, p. 794), in the Handbook of Industrial and Organizational Psychology, made this point in the following passage:

> Criterion measures are operational definitions of inferred constructs. Development of the construct is a matter of knowledge and wisdom; development of the corresponding measure is matter of technology. . . . . Only one point will be stressed.
>
> That point is the need to assess the construct validity of measures chosen as criteria. There are many approaches, such as systematic investigations of various sources of bias through correlational, experimental, or factor analytic methods. This is, admittedly, a prescription for hard investigative work, but a casual approach to the development and evaluation of criteria can only be expected to provide casual results. It is reasonable to suppose that prediction can be improved if criterion constructs are carefully thought through and the operational measures carefully studied for construct validity.

CRITERION MEASURES IN CONTROLLER SELECTION RESEARCH

Recent controller performance evaluation efforts within the FAA have included training performance measures, post-training job proficiency measures, and experimental computer-derived measures which may be applicable to both training evaluation and later job proficiency assessment. Training performance measures have consisted of academic phase grades, laboratory grades based on over-the-shoulder instructor evaluations, objective controller skills test scores, and pass-fail status at the end of Academy training. Post-training measures have included non-quantitative over-the-shoulder supervisor evaluations for diagnostic and administrative purposes, summary supervisor ratings for administrative purposes, quantitative supervisor evaluations for research purposes, post-academy attrition, and progression from low complexity job facilities to higher complexity job facilities. The same training measures have been used for both administrative and research purposes, while separate post-training evaluations have been collected for administrative and research purposes.

Performance Measurement During Training

Training level controller performance measurement procedures in the FAA (described in Chapters 9, 10, 11, and 12) generally represent state of the art methods. The written tests (academic exams and controller skills tests) have been subjected to considerably psychometric analysis at the individual item level and normative analyses have been performed across a number of samples. The controller skills tests were developed specifically to provide more objective measures of the application of knowledge and procedures taught during the laboratory phases

of training than were otherwise possible, considering the subjective nature of instructor evaluations of laboratory problem performance. Although the subjective instructor evaluations have been retained in the laboratory performance evaluation system, considerable effort has been put forth to standardize the problem scoring systems, in the hope that this would reduce some of the subjectivity in the scoring process. Further attempts have been made to minimize the effects of potential instructor bias by requiring that trainees have different instructors for different problems; thus the overall laboratory grade reflects evaluations by multiple instructors. Reliabilities of the individual instructor problem evaluations have been estimated to be about .40; however, when combined over six problems the reliability of the six problem composite was estimated to be about .80, which is generally regarded as acceptable in personnel research.

The work of Tucker (Chapter 12) on the development of dynamic paper-and-pencil simulations of air traffic controller problems for objective controller proficiency measurement represents a contribution that should be recognized as a significant advance of the state of the art in proficiency measurement. Although Tucker mentioned that the basic rationale and theoretical bases have been available for some time, dynamic written simulations have not been widely used for proficiency evaluation. Certainly Tucker's work has demonstrated the feasibility of applying the technique to the assessment of some aspects of controller activities and with additional research may be shown to be much more widely applicable than was initially considered possible.

A Radar Training Facility (RTF) was added to the FAA Academy at Oklahoma City in 1981 (See Chapter 9). The RTF has the capability of providing automotic computer-derived measures of student performance in moving computer controlled simulated air traffic, and one small-scale study (Boone, Van Buskirk, & Steen, 1980) was conducted to explore the feasibility of using objective computer-derived measures in place of subjective over-the-shoulder instructor evaluations for the measurement of student performance on the radar control laboratory problems. Considerably more research will be needed to validate the computer-derived measures, but the developments in this area certainly appear to be highly encouraging. The consideration of computer-derived measures is discussed further in a later section of this chapter.

The FAA has tended to utilize very elaborate weighting methods to determine overall scores and pass-fail status in the various phases of training; in these procedures, the more difficult problems and laboratory performance received higher weights (as described in Chapters 9 and 10). These weighting procedures have undergone extensive normative analyses, have received considerable scrutiny from technical experts, and have been shown to give higher weights to the components that differentiate between trainees and full performance level controllers. However, no data have been reported to validate the weighting schemes against actual job performance criteria.

As indicated in Chapter 13, very little research has addressed the relationship between training performance and later on-the-job performance. However, the research that has explicitly addressed this issue has indicated that instructor ratings and training grades correlate in the low .40 range with subsequent supervisor ratings of on-the-job performance. This is considered to be high for data of this kind.

The state of the art of training performance measurement in industry is difficult to determine, but does not appear to have much to contribute to the enhancement of controller performance measurement. Common practices range from instructor ratings to written quizzes to work sample tests to simulator performance evaluations, with considerable variation in psychometric quality within each type. In some instances the training performance measures have been examined critically in relation to such issues as possible bias, degree of differentiation among trainees, reliability, and relationship to later job performance. However, in many other instances, training performance measures have been accepted at face value.

Post-training Performance Measurement

Attrition. Post-training attrition has been a major criterion in controller selection research, undoubtedly reflecting concern with the training costs and time involved in bringing a controller trainee up to full performance level status. Attrition research in the FAA has been based on the assumption that most controller attrition is the result of failure by developmental trainees who lack ability to control air traffic effectively. Hence it has generally been believed that attrition could be reduced by increased selectivity at the point of entry, through the use of pre-employment aptitude tests.

The validity of the belief among FAA executives that aptitude deficiencies are causes of controller attrition has been supported by empirical studies which have found that instructor ratings of potential to become a full performance level controller were predictive of retention two years following initial training (Chiles and West, 1974) and that post-academy attrition dropped from 38% to 7% after the imposition of pass-fail standards in academy training, with a 30% academy failure rate (Boone, Chapter 9). However, Boone has also observed that a controller attrition rate of about 10% occurs for reasons unrelated to poor academy performance.

The only recent study which attempted to relate aptitude test performance directly to post-training attrition (Mies, Colmen, and Domenech, 1977) did not find a significant and consistent relationship between the two. A signficiant correlation of approximately .22 was reported between attrition and previous experience as measured by the Pre-Employment Experience Questionnaire (PEQ) described in Chapter 18.

In a recent review of the current literature in personnel selection, Tenopyr and Oeltjen (1982) noted that employee turnover (attrition) has not been a common criterion predicted by selection procedures. The general finding in turnover research has been that cognitive ability tests are not very predictive of turnover and that biographical information measures have typically yielded better results.

Muchinsky and Tuttle (1979) have recommended that turnover should be analyzed in terms of subgroups such as voluntary versus involuntary terminations, since on the empirical level studies that have made this distinction have found differential results for the two groups and on the theoretical level the explanatory mechanisms underlying the termina-tions are quite different.

Lofquist and Dawis (1969) have proposed a conceptual model of work adjustment in which turnover is a function of both job satisfactoriness and job satisfaction, where satisfactoriness is defined as the correspondence between individual abilities and the ability requirements of the job and satisfaction is defined as the correspondence between individual needs and the reinforcer system of the work environment. In this model turnover is predicted for individuals who do not achieve minimal levels of both satisfactoriness and satisfaction.

Most of the current turnover research outside the FAA has focused almost exclusively on issues related to the satisfaction component of the Lofquist and Dawis model (Mobley, Griffeth, Hand, and Meglino, 1979) while controller attrition research by the FAA has focused almost exclusively on the satisfactoriness component of the Lofquist and Dawis model. The difference in focus reflects the relative uniqueness of the air traffic controller situation, in which a major portion of the attrition occurs before the controller reaches full performance level status, primarily due to poor performance. It seems that any efforts to achieve further reductions in the controller attrition rate need to be based on an explicit differentation between attrition that is related to lack of job proficiency and attrition that occurs for other reasons.

Job progression. Job progression has been used as a criterion measure in only one recent controller validation study (Mies, Colmen, and Domenech, 1977, also reviewed in Chapter 18). These investigators ranked the four controller options, FSS, Terminal VFR, Terminal IFR, and EnRoute, in order of respective job complexity and recorded a progression score of 2 if a controller was currently assigned to either a more complex option or the same option as his initial assignment, and a progression score of 1 if the current assignment was a less complex option than the initial assignment. The progression score was based on a time period ranging from two to six years subsequent to initial assignment for various samples from classes that entered in different years.

Under the best of circumstances, job progression must be considered to be a rather crude and global indicator of individual performance. Unfortunately the progression criterion used by Mies, Colmen, and Domenech

270.

was further limited by the confounding facts that (1) the initial assignment could be in any of the four options, depending on availability of position openings rather than beginning at the lower rung of a natural progression with an initial assignment in the FSS option, and (2) later transfers to other options of either lower complexity or higher complexity were influenced by a number of personal choice and organizational factors unrelated to controller performance or capability.

The only relationship between progression and other criteria that could be computed in the Mies et al. study was with the supervisory ratings. The correlation between progression and supervisor's ratings was quite low (.10). In addition the progression criterion was not very predictable from the selection instruments: $r = .12$ for the test battery, $r = .08$ for the Personal Experience Questionnaire, and $r = -.06$ for the Occupational Knowledge Test.

Progression, in terms of promotion rates and level of management achieved, has been utilized as a global criterion in seveal major management selection studies at Standard Oil of New Jersey; Sears, Roebuck and Company; and American Telephone and Telegraph (Campbell, Dunnette, Lawler, and Weick, 1970). However, in those studies progression was defined essentially in terms of upward movement from lower entry level positions to higher level management positions, since downward movement was uncommon. In those studies progression was usually used in conjunction with supervisory rankings and salary history as a global indicator of managerial success.'

The "option complexity" job progression criterion appears to be of questionable utility as an indicator of controller effectiveness or success. It is recognized here that the purported rationale for including progression as a criterion was that downward movement represented inefficiency in terms of extra training cost as well as ineffective controller performance. However, combining it in an aggregate measure of controller success, where all of the other criterion components were purported to represent controller proficiency, seems to require an implicit assumption that movement to a lower complexity controller option occurs primarily due to lack of ability. If the administrative structure of initial assignment and movement from one option to another were to change, the progression criterion might become more meaningful, but at present it appears to be a highly contaminated measure of questionable validity and very limited predictive utility.

Subjective performance evaluations. Subjective performance evaluations conducted by the controller's immediate supervisor have been the most widely used measure of controller performance on the job. Over the years the FAA has placed more and more emphasis on over-the-shoulder proficiency ratings conducted "on-line," while observing the controller's actual job performance, as opposed to the common industry practice of periodically having supervisors rate, from recall, the typical performance of an employee over a period of several months to a year.

271.

The over-the-shoulder ratings have the advantage of focusing on observation of what the controller actually does during the limited duration work sample observed, but are correspondingly deficient in the respect that they do not provide an indication of the controller's day to day performance on the job. All too often, there may be a discrepancy between the two, as when motivation or attention lags. Provided that the traffic sample used for the check is sufficiently challenging, the over-the-shoulder rating should more properly be thought of as a maximum performance measure of the controller's capabilities, while the usual recall ratings are more representative measures of typical performance. The operational field performance evaluation system used by the FAA actually utilizes a system in which the detailed over-the-shoulder evaluations are used nonquantitatively for diagnostic and remedial training purposes; on a less frequent schedule, quantitative ratings are generated from a review of recent over-the-shoulder evaluations and any other relevant information that may be available (Henry, Kamrass, Orlansky, Rowan, String, and Reichenbach, 1975).

At first glance the over-the-shoulder evaluation procedures, especially when they incorporate carefully specified performance factors derived from job and task analyses, give the appearance of being more objective and thus more reliable than the traditional global ratings. However, Henry et al. have argued that, not only are the over-the-shoulder ratings subject to the same subjective biases as the global ratings, but when collected during real traffic control situations are subject to additional unreliability as a result of lack of comparability or standardization of the traffic samples being observed. No definitive comparison of the reliability of the two types of evaluation has been reported.

A related trend in controller evaluation has been the increasing use of rating forms with large numbers (30 to 50) of specific rating items, as opposed to forms composed of a smaller number of more global rating items. Research that has compared the two formats has found that the differences in reliability between the two have been slight and may actually favor the more global format (Cobb, 1967). Although it may seem that increased specificity should enhance objectivity and reliability, the increased number of items that need to be attended to at the same time by the rater easily exceeds normal human memory capacity and leaves the ratings subject to the considerable influence of selective recall and general impressions of the performance observed.

As noted in Chapter 13, current FAA staffing and field performance evaluation procedures preclude the opportunity routinely to obtain reliability estimates of the ratings, because of the unavailability of qualified alternate supervisors. Earlier reliability studies have produced varied findings, with reliability estimates of about .43 for typical field ratings and reliability estimates in the .65 to .75 range for carefully controlled studies using trained raters.

Supervisor field ratings of controller performance have been found to correlate in the .40 range with training performance and have been modestly predictable from pre-employment selection instruments, as shown by the respective correlations of .18, .17, and .15 with an experimental test battery, the Pre-employment Experience Questionnaire, and the Occupational Knowledge Test, in the Mies, Colmen, and Domenech (1977) study.

A review of the performance evaluation literature outside of the area of controller performance assessment reveals that current practices in the FAA are generally close to state of the art and that practices in industry offer little in the way of models for improvement.

Subjective judgmental indices of performance, usually some form of rating or ranking completed by the employee's immediate supervisor, have been by far the most common criteria used in industrial selection research. These subjective performance measures continue to be widely used in spite of the continuing lament regarding their subjectivity and questionable reliability and validity.

In view of the popularity of subjective rating measures and the continuing criticism of their recognized deficiencies, it should not be surprising that performance ratings have been subject to extensive research in an effort to develop improved evaluation methods and formats. A number of different rating formats have been developed (e.g., graphic ratings, behavioral expectation scales, forced choice ratings, behavioral check lists, alternation ranking, critical incidents) each with its own purported advantages and disadvantages. However, none has been shown to yield consistently superior results in empirical studies (Tenopyr and Oeltjen, 1982).

Behavioral expectation scales, developed using a retranslation method reported by Smith and Kendall (1963), have received a great deal of attention since they include several properties that should be expected to result in superior ratings. These include (1) a focus on actual job behaviors reported by supervisors, and (2) selection of behavioral anchors on which independent supervisors generally agree with respect to both the dimension of performance and the level of performance described.

Although the importance of reliability in criterion measures is widely recognized, studies which actually estimate and report reliabilities of the criterion measures used are in the minority. Recent meta-analytic research on the topic of validity generalization, which has attempted to aggregate results over multiple studies, has been forced to rely on assumed distributions of criterion reliabilities since the actual criterion reliability estaimtes for individual studies have rarely been reported. Based on reviews of available data, Schmidt and his colleagues (Schmidt and Hunter, 1977; Schmidt, Hunter, and Caplan, 1981) have developed assumed criterion reliability distributions with means of .60 to .70 for use in their research. They

273.

reported that these distribution are probably overestimates of the appropriate reliabilities since they generally were based on ratings provided at approximately the same point in time and thus did not reflect any changes in job performance over time.

Performance rating research provides little in the way of accepted guidelines for achieving high quality performance ratings, but three generally accepted conclusions suggest that: (1) scale development with respect to item selection and definition should receive very careful attention regardless of the type of rating format employed, (2) training of raters, including periodic retraining, in the use of the rating procedure and common rating errors to be avoided, tends to produce better quality ratings, and (3) special experimental ratings collected for criterion purposes generally yield less halo, greater variance, and higher reliability than ratings used for administrative purposes.

Keeping in mind the fact that research results reported in professional journals no doubt represent somewhat better practices than the typical practice in industry and that none of the research reported suggests any proven methods demonstrably superior to those currently used for controller performance evaluation, current controller rating procedures must be considered to represent near state of the art methods. However, one noticeable area where controller ratings have not been on a par with better industry practices is the failure to specify anchors or descriptions of controller behaviors in the rating items that would be representative of different performance levels (poor, standard, superior).

Continued efforts to refine rating practices might produce some improvements, but considered in the context of the finding of Buckley, O'Connor, and Beebe (1969) that the underlying behaviors that are assessed tend themselves to be somewhat unreliable and to vary from one time to the next, one has to recognize the futility of trying to solve the reliability problem through changes in the rating process alone.

Simulation and Computer-Derived Measures

There are two rather independent issues in the use of job simulations for performance evaluation purposes. The first involves the opportunity to manipulate and standardize the parameters and procedures used to create the simulated work exercises, and the second involves the possibility of using an automated system to record the actions and results of the actions of the subject. Although standardization of the simulation exercise is an essential and integral part of any simulation used for individual performance assessment, automated performance recording is neither essential nor integral to the simulation process and in many cases may be no more feasible than automated performance recording of the job itself.

Automated performance recording is actually an independent consideration which may be applicable to either the job itself or the simulation of the job or selected job activities. The critical considerations in the decision to adopt automated performance recording systems are (1) that the recording process not interfere with or change the job itself and (2) that the behavior of interest be amenable to automated recording. In addition, as with all performance measurement systems, the expected benefits from the system must be considered to be worth the cost of developing, installing, and operating the system. Computer based job environments and simulations are good candidates for automated performance recording since computer monitoring of system outputs can usually be achieved with relatively small additional cost, and without intruding on the work process.

The FAA has utilized air traffic control simulation research to evaluate the effect of workload variables and changes in control procedures on overall system performance for many years and has conducted several studies of the feasibility of using high fidelity job simulations to evaluate controller proficiency. In addition, the recently constructed computer simulation-based Radar Training Facility at the FAA Academy has both the capability of generating very realistic simulations of radar control activities and of providing objective computer-recorded system response measures. (See Chapter 9 for a more detailed description.) Much of the automated, computer based technology of the national air traffic control system itself appears to be readily amenable to automatic computer-derived performance recording.

Tucker, in Chapter 8, has noted that one of the current issues receiving high priority in the FAA is the development of an objective performance assessment system to replace the subjective over-the-shoulder evaluations of the current assessment system. Since the primary limiting factor in the simulation research to date has been the high cost of transporting controllers to a central simulation site or alternatively, of moving simulation facilities to various field locations, it would appear that objective field performance evaluation will necessitate the use of computer-derived measures of job performance in the operational traffic control system.

Computer-derived measures certainly offer advantages over subjective supervisory ratings, but are not without disadvantages either. Advantages include: (1) They are objective and thus possibly more reliable; (2) They can provide more precise and accurate measurement of some variables; and (3) They can provide performance assessment in terms of system outcome effects. Potential disadvantages include: (1) Some important job behaviors may not be assessable by system measures; (2) They provide no direct information about techniques and procedures used, although some of this information may be inferred from system results; (3) They allow for considerable possibility of contamination (built-in bias) due to system parameters that have nothing to do

275.

with individual controller proficiency; and (4) They may be extremely sensitive to variations in experience, job assignment, and non-standardization of work samples being recorded, which are factors that contribute to both bias and unreliability.

The experimental research using computer simluations for controller performance measurement has verified the feasibility of the approach by demonstrating that the simulation procedures were capable of producing a high fidelity representation of basic traffic control activities and that various system measures could be recorded automatically by the computer. Buckley, O'Connor, and Beebe (1969) further demonstrated that computer-derived performance measures reflected reliable individual differences among controllers in the research sample and that aggregate computer-derived performance measures correlated positively (.53) with observer's over-the-shoulder ratings of overall performance, for typical traffic densities.

Since the computer-derived measures are objective in the sense that they are not subject to the observation and evaluation errors inherent in supervisor's and observers' ratings, there is a tendency to assume that the computer-derived measures are also perfectly reliable or at least substantially more reliable than subjective ratings. The Buckley et al. study demonstrated that this assumption does not necessarily hold true; indeed considerable variability was found in the reliabilities of the various simulator measures across three traffic densities. The reliability of the computer-derived measures exceeded that of over-the-shoulder observer ratings only in the highest traffic density condition. This finding is particularly sobering when one considers that the traffic density in which the computer-derived measures were more reliable than over-the-shoulder evaluations was judged to be more difficult than that typically encountered under operational conditions. It should be noted that the reliabilities of both methods were in the marginally acceptable range of approximately .60 and .75, and that the acceptable range of approximately .60 and .75, and that the researchers found that two simulation runs of one hour in duration were required to achieve this level of reliability.

Much additional research will be needed to refine and validate simulation and computer-derived measures for operational use. Issues requiring resolution include selection and specification of measures that should be used, calibration of scoring procedures with respect to acceptable performance levels and tolerance ranges, determination of how the various measures should be combined to provide an index of overall performance, illumination of the meaning and implications of variations in the computer-derived measures in terms of controller behavior, determination of biasing factors in the system measures that need to be controlled experimentally or statistically, and determination of what should be done about relevant aspects of performance that cannot be assessed by the computer-derived measures. It should also be recognized that simulator and computer-derived measures will most likely be very sensitive to changes

in either hardware or work procedures of the controllers' job and thus will require frequent recalibration of the performance measurement system. Some of the complexities of developing and validating simulator-based performance evaluation systems are illustrated in considerable detail by Buckley, House, and Rood (1978) and Hennessy, Hockenberger, Barnebey, and Ureuls (1981).

SUMMARY AND CONCLUSIONS

Despite the limitations noted, the overall criterion development efforts in the FAA controller selection and validation research program have been impressive, especially when compared to other widely recognized selection programs. Historically some of the major recognized selection programs have been validated against rather limited criterion measures. For example the massive World War II U. S. Army-Air Force pilot selection program used pass-fail status in initial pilot training as the criterion and recent Air Force research on pilot selection was still using pass-fail status in undergraduate training as the criterion (Hunter and Thompson, 1978). The major programs in managerial selection (Campbell, Dunnette, Lawler, and Weick, 1970) have been based on the use of salary, promotion rate, management level, and ranking of overall effectiveness, as validation criteria.

The FAA controller selection research has included a variety of training and post-training criteria, both individually and in combination, to provide a broad indicator of controller success. Controller performance evaluation during training has received extensive research attention and has resulted in a strong training performance evaluation system. Post-training controller performance evaluation has been less impressive than evaluation during training, but nonetheless reflects substantial progress.

Post-academy attrition as a criterion developed out of a special combination of circumstances, that included concern with high training costs and a relatively high attrition rate that seemed to reflect primarily difficulties in learning to control traffic effectively. Reduction in attrition has been achieved through more stringent training pass-fail standards. Frequent field evaluations based on over-the-shoulder evaluations of controller proficiency have been standard administrative practice; these have been used mainly to diagnose remedial training needs and to assure minimal proficiency. Very specifically focused experimental ratings by supervisors have been collected for criterion purposes in selection validation studies in order to provide greater performance differentiation than is normally available with administrative ratings. A job progression criterion measure based on a hypothesized variation in job complexity among controller career options has also been used, although the theoretical and empirical support for this measure was considerably less than that for the other primary criterion measures.

277.

Research has been conducted to explore the feasibility of using simulators and objective computer-derived performance measures to assess controller performance in training and on the job. One of the goals of this research has been to develop objective controller performance measures to replace the subjective, over-the-shoulder performance ratings currently in use. More fruitful results would probably be achieved if simulator performance were cast in the framework of a work sample or performance test, with scoring systems developed to combine objective and subjective measures such that the advantages of each were retained in the fashion utilized in traditional performance tests (Adkins, 1947). For example, objective computer measures could be obtained for those variables that observers are unable to evaluate effectively, such as conflictions and delays, while over-the-shoulder evaluations could be used to assess performance factors that computers assess poorly, such as evaluation of control technique and adherence to recommended procedures. The present state of development in the application of simulation and computer-derived performance measurement indicates that it shows considerable promise but will be no simple task and will require substantial additional development and validation before meriting operational implementation.

The criteria utilized in controller selection studies have been quite useful in providing overall indications of controller success for the specific purpoes of the validation studies. Unfortunately the emphasis on global assessment of overall performance has not contributed much to the understanding of the nature of specific predictor-criterion relationships underlying the overall validity. Failure to focus on more specific relationships tends to make it much more difficult to discern which areas are likely to be good prospects for improvement or to anticipate the probable effects of expected changes in the job on the validity of various predictors.

It should be noted that throughout the work on controller performance measurement there has been no integrative theoretical perspective and that construct validation models of performance assessment will be required to make real progress in the future. This conclusion is consistent with Guion's (1976) suggestion that carefully thought through criterion concepts are necessary in order to improve prediction, and Hopkin's (1980) argument that significant improvements in controller performance measurement will require the formulation of controller activities in terms of psychological constructs rather than system functions.

PART IV.

RESEARCH LEADING TO THE 1981 ATC SELECTION BATTERY

Beginning around 1970, the FAA launched a major effort to develop a new test battery for the selection of Air Traffic Control Specialists (ATCS) to replace the Civil Service Commission (CSC) battery, which had been in operational use by the CSC and its successor, the Office of Personnel Management (OPM), since 1964. The major accomplishment in this effort was the development of two new tests, which are integral to the new battery that became operational in October, 1981, and added a major increment to the precision of ATCS selection when they were adopted. These are the Multiplex Controller Aptitude Test (MCAT) and the ATC Occupational Knowledge Test (OKT), both the work of John T. Dailey and Evan W. Pickrel, who, along with James O. Boone, were the principal architects of the research program and the new selection battery.

The first two chapters in Part IV report this new test development. They consist of Chapter 15, on the MCAT and Chapter 16, on the OKT, both by the test developers, Dailey and Pickrel. They are followed, in Chapter 17, by a definitive report on research on personality assessment of ATC applicants, by John J. Convey.

The following five chapters cover the principal research to identify an optimal new test battery for the ATCS selection task. These represent four validation studies based on trainee and controller samples, two normative studies of job applicants, and an analysis of the final battery in relation to the issues of adverse impact, validity, and fairness, for conformity with the Uniform Guidelines for Equal Employment Opportunity in selection. Chapter 18 summarizes two important validation studies carried out under contract with the FAA by Joseph G. Colmen and his associates at Education and Public Affairs (EPA), a private consulting company, and a major study by James O. Boone, at the FAA Civil Aeromedical Institute (CAMI). Together, these three studies represent the major developmental research that produced the new selection battery. Chapters 19 and 20 report two normative studies of ATCS job applicants which provided needed information on the performance of the applicant population on the CSC and the new experimental tests, normative data by sex and race-ethnic group, as well as for analysis of preference credit awarded to veterans by law. Chapter 21 presents an independent validation of the experimental battery in comparison with the CSC battery. And Chapter 22 presents information on adverse impact, validity, and fairness, which justify the adoption of the experimental battery under the uniform guidelines.

The final chapter in Part IV, Chapter, summarizes the research presented and the conclusions and recommendations of the FAA-OPM investigators responsible for the new battery. Chapters 18 through 23 were prepared by S. B. Sells, from summaries presented in the FAA report by Rock,

Dailey, Ozur, Boone, and Pickrel (1982), which documented this research program in preparation for operational adoption of the new battery.

280.

Chapter 15

DEVELOPMENT OF THE MULTIPLEX CONTROLLER APTITUDE TEST

John T. Dailey and Evan W. Pickrel

INTRODUCTION

This chapter describes the creation of the Multiplex Controller Aptitude Test (MCAT), a new measure to screen applicants for Federal Aviation Administration Air Traffic Controller positions (Dailey and Pickrel, 1977). Its content includes both items that measure the traditional types of aptitudes, such as arithmetic reasoning and visualization, that are included in many tests employed by the Office of Personnel Management, and new job sample items derived from analysis of the controller activity. The job sample items utilize figures to show air traffic on a simulated radar screen, and require identification of potential conflicts between aircraft in that simulated traffic. All test questions are presented in an air traffic control setting, providing a job-related appearance that gives the test high face validity not found in the predecessor selection battery which the MCAT was designed to supersede.

The format for item sequencing in the MCAT also departs from the traditional aptitude battery design practice of clustering items into homogeneous subgroups. Instead, the MCAT items alternate from one type to another and spiral to increasing levels of difficulty. This mode of presentation is found only in a few current tests, most prominently in the Stanford-Binet. A result shown statistically is that non-confliction items, included in the former battery for measurement of aptitudes show unexpectedly high commonality with the new air traffic items that require detection of impending conflictions. The resulting measures have consistently produced higher validity coefficients than those obtained with the earlier battery.

In 1970, the Federal Aviation Administration initiated efforts to develop new tests to replace the Civil Service Commission (CSC) selection battery that was adopted for operational use in 1964. Among the numerous leads investigated, the most promising idea was adapted from a technique developed by Buckley (Buckley, O'Connor, and Beebe, 1970; Buckley and Beebe, 1970) entitled the Controller Decision Evaluation technique (CODE). This involved a motion picture film that presented simulated air traffic in real time as seen in a controller's display scope and was developed by Buckley to evaluate proposed air traffic displays by measuring the facility with which a specialist could control the simulated traffic shown on the film.

This idea was adapted by the present authors for controller selection and, in the course of development, CODE became MCAT, after a series of transformations from (1) the original, unstructured, free response film test to (2) a structured film-slide version of CODE, in which conflict questions were superimposed on the film by means of slides, to (3) a film-slide MCAT, in which aptitude questions were added to the conflict questions in the structured CODE test, to (4) an all slide

version, to (5) a paper and pencil version of the slide test. Early versions of this test were found by Milne and Colmen (1972, see Chapter 18, Part I) to add significantly to the validity of an aptitude test composite in the prediction of on-the-job success of air traffic controllers. However, the equipment and logistics required for its administration were generally unsuited for use in Civil Service Commission testing situations and the search was continued to develop a more acceptable format while preserving the predictive validity of the test.

TEST DEVELOPMENT

The CODE Test

Unstructured film version. The original Controller Decision Evaluation technique (CODE), by Buckley, O'Connor, and Beebe (1970), required observation of simulated air traffic situations as they unfolded in a film of a radar scope. Test versions of CODE were adapted in which observers were required to predict and record potential conflicts as soon as they detected them; the initial CODE tests involved an unstructured, free response mode.

During experimental administration and study of three film versions, it was found that identification of potential conflicts was easier, quicker, and more accurate (a) after observers adjusted to the scope and its targets, (b) when only a few targets were in near-confliction, and (c) when their rate of closure was slow. By contrast, items were more difficult when the observers were (a) first exposed to the scope and its targets, (b) when multiple targets were in near confliction, and (c) when the rate of closure among targets was rapid. False-positive conflict items did not discriminate between high and low performers, and for true conflicts, the discriminating power of the items was generally greater as the lead time available increased. Journeyman controllers identified potential conflicts sooner than developmental trainees, and they identified almost all potential conflicts that became real conflicts. Developmental controllers generally missed a number of real conflicts, were slower than journeymen in calling them out, and frequently did not attend to aircraft altitude separations. However, the range of test performance among developmental controllers was great, and some performed as well as journeymen. Some confliction items in these films were too easy and some had a negative relation to the developmental-journeyman criterion dichotomy.

A limitation of the free response, initial versions of the test was that they allowed the observer too much idle time, since the traffic presented was light and the aircraft were widely separated. In this sense, the setting resembled a common air traffic situation. However, an extended amount of testing time was required to present only a small number of confliction items, and more efficient use of the examinees' time was desired.

Structured, film-slide version. In the process of test development, an initial change involved the introduction of structured items, this was

accomplished by the use of slides to present questions on the screen above the film. Mean response times to react to each conflict were determined in the free response versions for both developmental and journeyman controllers and the mean times for journeyman were employed as a basis for selecting the aircraft positions for presentation of confliction questions in the structured versions. This had the effect of maximizing the discriminating power of each item. Two-choice conflict items, asking whether a pair of aircraft would conflict, were presented for 30 seconds when unexpected changes occurred among targets, such as when new aircraft entered the picture. Four-choice conflict items, asking which pair would conflict if a confliction should occur, were presented for 45 seconds when traffic changes were slow. A "None of these" response was introduced to permit inclusion of non-conflictions that might be predicted to be conflictions, and this was used as the fourth alternative in all four-choice confliction items.

The items were assembled in the described test format, but still, they utilized less than half of the available testing time. In addition, the large amount of idle time between questions provided opportunities for observers to change their answers to earlier questions regarding potential conflicts as the aircraft approached each other and the correct answers became more obvious. In the free response film version, this problem had been controlled by a requirement for entries from a coded clock onto the computer-scored answer sheet whenever potential conflicts were reported. (In the structured version, it was found that the presentation time of a new item every 45 seconds would keep examinees so busy that they would have no opportunity to make such changes; therefore, more test items were needed to make this control procedure work.)

Transition to the MCAT

Initial film-slide version. The test films presented simulated traffic moving across a radar scope, and a table that included detailed data on each aircraft, identifying the target, its altitude, speed, and route on the scope. The scope used lines to represent airways or highways in the sky, with alphabetical identifiers of starting, ending, and intersecting points on the airways. The orientation of the scope was with North at the top and a mileage scale was provided at the bottom. Ample information was thus presented to enable the preparation of a variety of additional questions related to controller activity. It appeared possible that in addition to the confliction items, most of the factors included in the then current Civil Service Commission aptitude battery for controller selection could be measured within this simulated air traffic setting.

Items were written utilizing this available detailed information to measure such aptitudes as direction following, table reading, interpretation of data, spatial visualization and orientation, estimation of distances and relative target movements, and arithmetic. Some of the items included were very simple, and others were written in a multi-factor format to increase their level of difficulty. For example, the directions

283.

for this type of test required instruction on how to read the table; therefore, initial table reading questions were included that were of the very easy, instructional type. The complex items, such as estimating the flight time (in minutes) between two aircraft at a given moment, required awareness of distances across the scope, reading the table to determine the speeds of aircraft, and mathematical computation to determine their·rates of closure (all related to horizontal separation.

The relations between item types and total test homogeneity were calculated, and this ratio was used to determine the number of items per type to include in the test. A result, for example, was the inclusion in the test of twice as many target time-distance separation items as compass heading items. The order of placement of conflict items had already been established by using the mean response time of journeyman controllers when targets were at certain locations. New aptitude test items were interspersed in the remaining positions on the film, alternating from one type to another and spiraling to increasing levels of difficulty from the beginning to the end of the test. Three alternate forms of this film-slide test were prepared, entitled MCAT 4, 6, and 7.

Versions of the free response CODE test, the structured film-slide MCAT, and other selection measures were administered to 109 students at the US Navy Air Traffic Controller Training School, Memphis, during the week of August 18-22, 1975. Grades of these students were obtained as they progressed through classroom and laboratory training and passed or failed the course.

Table 1 compares the correlations of the free response CODE test, the film-slide MCAT, and the Occupational Knowledge Test (OKT) (See Chapter 16), the Navy General Classification Test (GCT), and a Navy Arithmetic Reasoning test with course grades in controller training. The highest correlations were obtained by the OKT, a measure of ATC-relevant knowledge reflecting job-related experience and training, and the MCAT and GCT were next in magnitude. The correlations for MCAT exceeded those for the free response CODE test. In general, these results were regarded as highly encouraging and further development of the MCAT proceeded.

Development of the slide-only version. The next step in test development was to convert the film-sli'e version, which presented moving targets, to a "slide only" presentation, which in essence captured pictures of the scope with targets at the precise position they were in when each question appeared on the film screen. Pacing of the slide presentation was the same as in the film-slide version; as a result, the amount of target movement from question to question was unchanged. This format was more convenient for test administrators, since they no longer had to cope with film projection problems, and the stationary targets presented by slide seemed easier to read than the moving targets presented by film.

Table 1

Correlations of selection test variables and school grades

USN Air Traffic Control Training School

(109 persons tested August 18-22, 1975)

School Courses	Correlations of Tests with School Course Grades				
	Free Response CODE Test	Film-Slide MCAT	OKT	GCT	Arithmetic Reasoning
Laboratory Flight Plans					
VFR	.05	.17	.23	.10	.05
IFR	.18	.26	.22	.19	.26
Stopover Composite	.09	.20	.28	.26	.09
Performance Run	.19	.22	.22	.25	.18
Laboratory Control Tower Training					
Basic	.16	.20	.28	.19	.21
Intermediate	.19	.11	.07	.17	.09
Advanced	.26	.17	.27	.19	.08
Laboratory Radar Training Air Surveillance					
Week-1	-.02	.10	.11	.17	.10
Week-2	-.08	.04	.21	.09	.12
Precision Approach					
Week-3	.08	.21	.13	.14	.08
Week-4	.16	.04	.09	.11	.25
Course Average	.22	.27	.43	.34	.23

285.

Persons taking the experimental version in this slide format did grumble that the test seemed to take control over their time, as each item had to be completed quickly before it left the screen and a new item appeared. However, this task characteristic has some commonality with controller situations in the real work, as traffic movement tends to control their involvement and pace of work. A major merit of this version was that it could easily be converted to a paper and pencil format that would meet Civil Service Commission requirements for use in their highly decentralized field testing situations.

The slide-only version of the MCAT, three forms of the film-slide version, and two forms of the OKT were administered to 461 students at the USAF Air Traffic Controller Technical School, Keesler Air Force Base, Mississippi, along with other selection measures, during the week of October 20-24, 1975. Grades of these students were obtained as they progressed through classroom and laboratory training and passed or failed the course.

Results for the Air Force controller trainees are shown in Table 2. Here again the OKT (Form I) had the highest overall correlations with the training grades. This test also correlated most highly with Block IV and indeed was the only one of the six predictors that showed a substantial correlation with that measure. For the first three blocks, MCAT 606 AS, the slide-only version, was most consistently correlated with training success, slightly better than OKT I and generally above the three film-slide version forms. These results gave further encouragement to the development effort.

The paper and pencil version. The next step in test development was to print the slide versions in paper and pencil format. The slide version and combination film-slide version both permitted time control at the individual item level, allowing thirty seconds for response to two-choice items and forty-five seconds for response to four-choice items. One paper and pencil version allowed uninterrupted work throughout the testing time, with announcements when fifteen minutes elapsed and when only five minutes remained to work on the test. A segmented paper and pencil version allowed five minutes for response to each cluster of items, when they were all of the two-choice type. Three parallel forms of the written test were prepared; these were MCAT 406 with 41 items, 606 with 43 items, and 706 with 53 items. The time limits were 25 minutes, in addition to directions, for MCAT 406 and 606, and 30 minutes, in addition to directions, for MCAT 706. Each form has since been extended to 55 items. Three additional forms, 407, 607, and 707, were later constructed: MCAT 707 contained 23 conflict and 32 aptitude items; and MCAT 707 co - tained 23 conflict and 32 aptitude items; the time limits were 35 minutes (in addition to instructions) for each of these tests.

Estimation of reliability. The questions presented in the film-slide version of the MCAT, Forms 4, 6, and 7, remained unchanged through several versions of the test, although the media employed in presentation of the
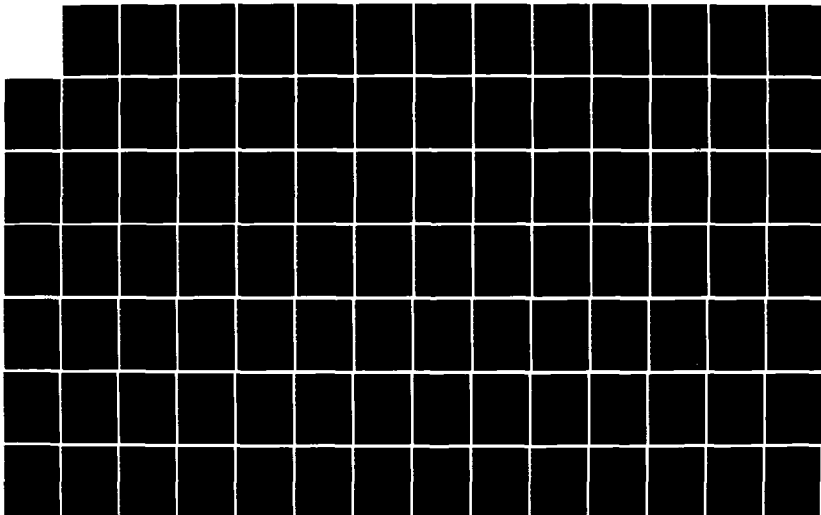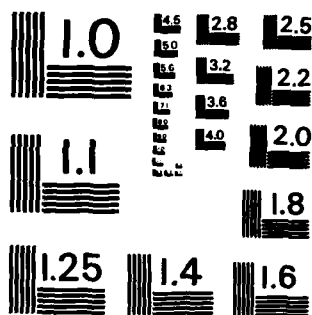
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS – 1963 – A

Table 2

Correlations of 4 versions of the MCAT and 2 forms of the OKT
with school grades. USAF Air Traffic Control Training School,
Keesler AFB, Miss.   October, 1975.   N=461

Predictor Tests	Training School Course Grades			
	Block I	Block II	Block III	Block IV
MCAT 606 AS (slide only)	.35	.21	.33	.06
MCAT 606 FS (film-slide, 43 item)	.15	.15	.16	.18
MCAT 706 FS (film-slide, 53 item)	.20	.21	.17	.07
MCAT 406 FS (film-slide, 41 item)	.19	.30	.27	.01
OKT - I	.42	.15	.32	.68
OKT - II	-.01	.09	.39	.01

items changed from film to film-slide, to slide, to paper and pencil. The changes in the media of item presentation had some effect on individual test performance, but all forms were qualitatively similar. The correlations of these tests with the same items presented in different media provide a minimum estimate of the test-retest reliability of the MCAT, when administered to members of similar populations.

These forms were administered to students of the USAF Air Traffic Controller Technical School in all phases of training and to a sample of 40 non-controller high school and college level FAA employees, groups whose abilities approximate those of the general population of applicants that take the Civil Service Commission tests. They were also administered to entering FAA Air Traffic Controller Academy students in the new centralized Air Traffic Controller training program, who were selected in the order of highest scores on the Civil Service Register of controller applicants. Table 3 presents distribution statistics for these groups on the various forms of the test. The range of test scores tended to be restricted for most of the ATC Academy classes, compared to the other groups.

Correlations between forms were lowest, .31 to .50, for the ATC Academy students, for whom the range of scores was most restricted. They ranged from .61 to .66 for the USAF ATC students and from .87 to .90 for the small sample of non-controller FAA employees, whose abilities more nearly approximated the applicant population for whom the test was designed. These coefficients indicate that the test can be expected to provide reliable measures of performance under well controlled conditions; by combining the scores for two forms an even more reliable measure would be assured for operational testing.

Validity of the Paper and Pencil Version

The EPA study. Validation of the paper and pencil MCAT was carried out by Mies, Colmen, and Domenech (1977) under an FAA contract with Education and Public Affairs (EPA), a private research firm (See also Chapter 18, Part II). This study utilized 2 parallel forms of the test -- MCAT 606 (with 43 items: 18 conflict and 25 aptitude) and MCAT 706 (with 53 items: 22 conflict and 31 aptitude) as part of a prediction battery that included 6 other cognitive tests (Directional Headings, Dial Reading, Arithmetic Reasoning, General Information, OKT, and the then current 5-part Civil Service Commission Battery) and 4 other instruments (Pre-employment Experience Questionnaire, Concept Adjective Test, Biographical Inventory, and 16 Personality Factor Questionnaire).

This battery was administered to 610 trainees in three 1976 FAA Academy classes, who were in the Terminal and EnRoute training options, to 488 ATSC's employed in 1973-74 (presumed to be at the Developmental level), and to 491 ATSC's employed in 1969-70 (presumed to be at the Journeyman level); all persons tested were volunteers, not in supervisory positions, and under 31 years of age when appointed.

288.

Table 3

Multiplex controller aptitude test

Distribution statistics for various populations

| | FAA ATC Academy Classes | | | | USAF ATC School Students | | | | FAA Employees | |
| | January 1976 | | March 1976 | | | | | | College | High School |
Tests	Enroute N=91	Terminal N=106	Enroute N=87	Terminal N=106	N=316	N=141	N=289	N=62	N=18	N=22
MCAT 606 (43 Items)										
FS Mean					28.81				30.73	17.45
SD					5.32				6.15	3.44
AS Mean						27.35				
SD						6.08				
A Mean	35.16	35.24								
SD	3.90	3.95								
B Mean			32.22	32.12						
SD			5.53	4.71						
MCAT 706 (53 Items)										
FS Mean							30.51		36.23	17.23
SD							5.83		8.49	4.32
AS Mean	39.96	37.11								
SD	4.40	8.68								
A Mean			40.20	39.15						
SD			7.48	7.41						
B Mean										
SD										
MCAT 406 (41 Items)										
FS Mean								20.47	30.50	13.89
SD								4.87	4.08	4.96
A Mean										
SD										

FS: Film-slide version, with time controlled for working each individual item.
AS: Slide version, with time controlled for working each individual item.
A: Paper-and-pencil version, with time controlled for working total test.
B: Paper-and-pencil version, with time controlled for working each five-minute group of items.

289.

For the purpose of test validation, four hierarchically related criteria were adopted. These were (1) performance during the initial, formal training period at the FAA Academy, (2) on-the-job performance at the developmental or journeyman level, based on supervisor ratings, (3) progression or mobility within the ATC system, scored by comparing the complexity level of current assignment with that of initial assignment, and (4) attrition, defined by whether or not the individual had separated from FAA as an ATCS. An aggregate assessment, on a 5-point scale, was also constructed, which combined the four individual measures. Information on progression and attrition and test scores on file from earlier testing were retrieved for analysis along with the scores on the current battery and supervisor ratings obtained for the sample tested.

The trainees enrolled in the initial, formal training program in the FAA Academy represent a highly select group, for their scores on the Civil Service Commission test battery were all above the qualifying level, and this selectivity tended to increase as those who failed to learn or perform adequately were separated over time. This selection process continues as the trainees progress up the ATCS career ladder and are evaluated for satisfactory performance on the job and advancement into the more demanding and higher paid positions within the ATC system. The range of scores on these selection tests can therefore be expected to be greatest for the group when they first enter formal training, and to become restricted with career progression. Since restriction of range is known to have a direct effect on the size of the validity coefficients, the validities should be highest for the group in initial training and somewhat lower for the survivors of the initial group as they progress up the career ladder.

Success in training. Validation against training criteria was performed for the sample of 610 trainees admitted to the Academy in 1976. The MCAT was scored separately for conflicts and aptitudes, and the criterion of training success, for those enrolled in the EnRoute options, was based on scores received on four ATC laboratory problems and scores on a Controller Skills Test. For those enrolled in the Terminal course, performance was based only on the ATC laboratory problems. These scores require students to demonstrate operational application of academic knowledge.

The results reported in Table 4 (see upper section – Training Scores) show the validity coefficients of the MCAT (sum of both forms) conflict and aptitude scores, the Arithmetic Reasoning, Directional Headings, and Dial Reading tests, and the multiple correlation of this selected battery in relation to composite measures of training performance. It is apparent that the MCAT validities, for both training options, exceeded those of the other predictors. More important, however, the other tests increased the joint prediction of the two MCAT scores by only .017 for Terminal trainees, .012 for EnRoute trainees, and .008 for both options combined.

Performance on the job. Test validation in the operational situation was accomplished by correlation of test scores with supervisory ratings and career progression scores. On-the-job performance of the 1973-74 and

290.

Table 4

Correlations between selection tests and various criteria

| | | | Selection Tests | | | | |
| | | | MCAT | | | | |
Criteria	Number of Cases	Arithmetic Reasoning	Conflicts	Aptitudes	Directional Headings	Dial Reading	Selected Battery
Training Scores[a]							
Terminal Option	301-310	.136*	.321**	.256**	.191**	.288**	.338**
EnRoute Option	257-263	.276**	.434**	.402**	.218**	.321**	.473**
Both	558-573	.202**	.370**	.323**	.202**	.267**	.395**
Supervisory Rating[b]							
Flight Service Station	195-245		.254**	.276**		.204*	.293**
Terminal VFR	169-190		.227**	.220**	.221**		.289**
IFR	188-188		.154**	.048			.166
EnRoute	181-241		.075	.125		.026	.128
All Options	733-833		.156**	.151**	.123**		.176**
ATC Progression[c]							
Flight Service Station	157-159	.148	.118	.231**		.075	.232**
Terminal VFR	191-193	.107		.133	.120		.178
IFR	179-179	.058			.170		
EnRoute	199-200		.298**	.285**		.280**	.357**
All Options	727-731		.094*	.117**		.074	.119*
Aggregate Criterion[d]							
Flight Service Station	196-249		.220**	.262**		.197**	.272**
Terminal VFR	479-518		.254**	.195**	.179**	.157**	.275**
IFR	499-623		.235**	.178**	.146**	.134**	.246**
EnRoute	455-514		.253**	.264**	.158**	.166**	.287**
All Options	1309-1603		.238**	.223**	.174**	.149**	.265**

*p<.05   **p<.01

Data are from Mies, J. M., Colmen, J. G., and Domenech, O. (1977).
a   Table XI. 1
b   Table XII. 2
c   Table XIII. 2
d   Table XV. 2

291.

1969-70 ATSC's was measured by confidential job-task assessments prepared by each employee's supervisor. Progression data are discussed in the following section.

Correlations between selection tests and supervisory ratings are presented in the second section of Table 4. Arithmetic Reasoning scores were not available for these groups. The correlation between MCAT conflict and aptitude scores and supervisor ratings was lower than for training measures, but nevertheless significant for the FSS and Terminal VFR options and for all options combined. The correlation between MCAT conflict scores and Terminal IFR supervisor ratings was also significant. The MCAT conflict and aptitude scores were prominent predictors in most of the regression equations that predicted job performance success, although neither predicted at signficant levels of confidence for the En-Route option. The Directional Headings test was significantly correlated with supervisor ratings only for the VFR option and for all options combined, and the Dial Reading test, for FSS only.

ATC Progression. ATC Progression was measured by comparison of the ATC option to which the specialist was assigned initially, when hired, to the option assigned on January 1, 1976. Progression from FSS to Terminal VFR and IFR and EnRoute options was considered as representing increasing levels of complexity. A progression of "high" was assigned within this hierarchy when a specialist was in an option of a complexity level the same as or higher than the initial option assigned. A progression value of "low" was assigned with a specialist was in an option of a lower complexity level than the option to which initially assigned. Correlations between selection tests and progression scores are presented in the third section of Table 4. Here, the MCAT conflict and aptitude scores were the best predictors in the EnRoute option and were significant for all options combined. MCAT aptitude was also significant for FSS. The total battery prediction was significant for VFR, EnRoute, and all options combined.

Aggregate Criterion. An aggregate criterion of ATC "success" was constructed from combinations of the four individual criteria (training, on-the-job performance progression, and attrition) and provided a five point scale value for ATC success. The Arithmetic Reasoning test was significantly related only to the progression criterion, and was dropped from further consideration in the selected battery. The Directional Headings test was dropped for FSS, since it failed to enter the regression for supervisory assessment. Correlations between the selection tests and the aggregate criterion are presented in the lower section of Table 4. All correlations were significant for each of the four options and for all options combined. The Dial Reading test did not enter the regression for the VFR option and entered last with a negative "b" weight for IFR and for all options combined. The integer weights assigned to tests are summarized in Table 5. It is evidence that MCAT (conflict and aptitude scores) was the major factor in the validities derived from the multiple regression analysis.

Table 5

Integer weights assigned to tests

by ATC option and all options combined

ATC Options	MCAT Conflicts	MCAT Aptitudes	Directional Headings	Dial Reading
FSS	18	40	(0)	7
VFR	55	10	21	(0)
IFR	49	8	13	(0)
ARTCC	33	34	7	5
All Options	37	21	15	(0)

Note. Data are from Mies, J.M., Colmen, J.G. and Domenech, O.,
May 1977, Table XV.3

293.

Further psychometric development. The development of parallel forms, calibration of length, time limits, and scoring of the MCAT, and results for sex and race groups are described in detail in Chapters 18 through 22, and in Rock, Dailey, Ozur, Boone, and Pickrel (1982). This report also describes the results for administration of one, two, and three forms simultaneously, which led to the decision to administer two forms for operational testing. (using the sum of total scores for the two forms).

The Boone study. Further evidence concerning the validity of the MCAT was provided by Boone (1979a) in research initiated at the FAA Civil Aeromedical Institute (CAMI) to compare a battery of experimental controller selection tests, including the MCAT, with the then current Civil Service Commission Test Battery in relation to success in Academy training. The study involved 1827 students who entered training between July 1976 and May 1978; an additional 1181 students were excluded because of incomplete data on the CSC tests, the laboratory criterion scores, or other measures, or because they did not volunteer for the experimental tests.

The MCAT measure in this study was a standard score that enabled combination of scores from several different versions of the test (Forms 606A, 606B, 706A, 706B, 607, 707) that were administered to different classes and that varied in the number and mix of questions (aptitude and conflict) and in the length of time allowed for the test. The ATC laboratory average score, standardized separately for the Terminal and EnRoute trainees (referred to as ZLab) was utilized as the measure of training success.

A significant feature of this study was that correlations with the criterion, ZLab, were corrected for restriction of range; the unrestricted correlations approximate the magnitudes that might be expected if the test were administered to a representative sample of applicants and if all persons tested were admitted for training. For the total sample, the obtained (restricted) correlation of MCAT total score (conflict + aptitude) with ZLab was .277, and the unrestricted correlation was .531. Both coefficients exceeded those of all other experimental tests in the study (Directional Headings and Dial and Table Reading) and of all of the tests in the CSC Battery. For a recommended replacement battery, consisting of CSC 24 (Arithmetic Reasoning), CSC 157 (Abstract Reasoning), and the MCAT, the multiple correlation with ZLab was .545 and the weights assigned to the three tests, based on beta weights in the regression equation, were 1, 2, and 4, respectively.

Restriction of Range. Adequate correction for restriction of range should accompany any comparison of experimental and operational tests. Efforts were made here to accomplish this, but no really adequate correction seemed possible because of several practical limitations of the data. Two of the operational Civil Service tests had pronounced negative skew that severely distorted any estimates of unrestricted validity. It is also hard to compare unrestricted validity estimates of tests administered under operational and experimental conditions. Variations in

294.

Table 6

Restricted Correlation Regression Analyses
New Test Battery and OKT
By Race, Sex, and Total Group

		Test Battery			Test Battery plus OKT	
	N	R	R^2		R	R^2
Total Sample	592	.416	.173		.477	.228
White	545	.368	.135		.440	.194
Men	515	.406	.165		.462	.214
Women	67	.523	.273		.613	.376
Black	39	.689	.475		.759	.579

administration sometimes distort the relations between test validity and
test variance, especially with highly speeded tests with short time limits.
There is also the question as to whether the test-criterion regression
is linear at the extremes when cutting scores are very high. With a very
high degree of selection the restricted validities will be very low and
the corrected validities will be extremely unstable. As a result of the
above considerations the validity coefficients have been often presented
both corrected and uncorrected. As a result no really precise compari-
sons of the validities of the operational and experimental tests has been
possible.

The OAM study. The final study was conducted by the FAA Office of
Aviation Medicine (OAM) under the supervision of the present authors and
is summarized by Rock, Dailey, Ozur, Boone, Pickrel (1982) and in Chapter
21, in this book. This study involved 953 Academy trainees who entered
training during the period June through December, 1978. The predictors
were MCAT (Forms 4o6e, 4e6o, 6e7o, 6o7e, 7e4o, and 7o4e), CSC 24 (Arith-
metic Reasoning), CSC 157 (Abstract Reasoning), and OKT 101B, 101C, and
102. Two parallel forms of MCAT were administered to each student. The
criterion measure of training success was pass-fail in the Academy.

In this study, corrections for restriction of range were not made.
However, the restricted correlations of MCAT with the pass-fail criterion
were somewhat higher than in the CAMI study, where the ZLab score was
used as criterion. For the first MCAT administered to each trainee (of
the two that were administered), the correlation was .348 (compared to
.277 in the Boone study). The correlation for the second MCAT admini-
stered was .405. Research with this test has shown a slight increase in
correlation for a second form of MCAT, as in these results, but no further
increase when a third form was administered.

Table 6 shows the multiple correlation (restricted) of the new bat-
tery with pass-fail, and for the battery and OKT (for which the correla-
tion was .429) for the total sample and by race and sex groups.

SUMMARY

The MCAT was developed as a new test for preemployment screening of
applicants for positions as air traffic controllers. The basic idea for
this test arose from the Controller Decision Evaluation (CODE) technique
developed by Buckley (Buckley, O'Connor, and Beebe, 1970) which involved
a motion picture film that simulated the display of air traffic as seen
in a radar scope. Three CODE films were adapted to testing by requiring
the observer to report potential conflictions of pairs of aircraft in
the display. Test development progressed from the initial free-response
version, to a structured film-slide version of CODE and then to a film-
slide MCAT, in which aptitude questions were interspersed among the con-
fliction questions. Subsequently, the film-slide version gave way to a
slide-only version and finally, the paper and pencil version. Further
development involved reliability studies, validation in 3 independent

296.

studies, and development of alternate forms for use in operational
testing. The MCAT represents a major advance in the technology of
selection testing and a major improvement in the selection of air
traffic controllers in respect to content, format, reliability, and
validity.

297.

298.

Chapter 16

DEVELOPMENT OF THE AIR TRAFFIC CONTROLLER
OCCUPATIONAL KNOWLEDGE TEST

John T. Dailey and Evan W. Pickrel

Subject matter knowledge or information tests designed for personnel selection include some of the oldest as well as some of the newest selection tests (Boone, 1979a). The famous Army Alpha test for enlisted men in World War I covered such information areas as animals and birds, physical science, biological science, farming, ranching, mechanics, social science, electricity, medicine, games, business, foods, airplanes, sports, guns, law, literature, and art. During World War II, information tests were widely used for selection, especially in the fields of aviation and electronics. Since then the Armed Forces have made extensive use of information tests in their selection and classification programs. The areas covered by these tests include mechanics, tools, physics, electricity, shop, radio, automotive, electronics, and general information.

One important application of information tests for selection is the Electronics Technician Selection Test of the U. S. Navy (Cobb, 1968). This test was originally developed during World War II for use in recruiting electronics personnel with prior experience in that field. The test consists of five basic types of subject-matter items, namely mathematics, general science, shop practice, electricity, and radio. When added to the regular Navy aptitude tests, it was found to increase the validity of the aptitude battery in the prediction of final grades in Electronics Technician School from .66 to .77. It was also found to have very good validity for other technical training schools including Radarman, Radioman, Sonarman, Aviation Ordinanceman, Electrician's Mate, Tradesman (Training Devices), and Aviation Electronics Technician. The Electronics Technician Selection Test is very similar in concept to the Air Traffic Controller Occupational Knowledge Test described in this chapter.

In 1960, Project Talent (Note 2), in the first aptitude census of 500,000 high school students, employed a wide variety of information tests and these tests have been found to have useful validity in relation to a number of important criteria (Dailey and Shaycoft, 1961). It was found that the 43 tests in the Project Talent test battery fell into only seven basic types. One of these was English Achievement, one was Spatial Reasoning, one was Perceptual Speed, and the other four were all represented by information tests in the areas of Verbal Information, Mathematics, Science and Technology, and Hunting-Fishing-Mechanics-Farming.

It can be seen that there is much precedent for the use of subject-matter information tests together with the traditional types of aptitude tests in selection programs. The Civil Service Commission, of course, has always used professional information as well as aptitude in its programs.

Experimental Test Development

The initial ATC Occupational Knowledge Test was designed to measure the acceptability of an applicant's claimed experience for qualification at the GS-9 level or above. There was such a wide variety among applicants' claims of experience that they were difficult to evaluate, and it was hoped that the OKT would provide scores to verify the acceptability of claimed experience. The fir t step in test development was a two-week workshop in June 1970, in which eighteen journeyman air traffic control specialists from Terminals, Centers, and Flight Service Stations participated. The workshop was organized to teach these specialists the specifications for the test and to help them to write items under the supervision of representatives of FAA Headquarters. This committee examined in detail the types of qualifying experience that were given credit, and determinated the specifications for the test, to permit its application to the verification of claimed experience. Then over 300 items were written, reviewed, and edited by the Examination and Certification Section of the ATC Academy, Oklahoma City. Two test booklets were created, Form 15 containing 150 items, and Form 16 containing 160 items.

These forms were administered to two classes of students on the day they reported for training at the ATC Academy, and to samples of GS-7 and GS-9 Air Traffic Control Specialists in Flight Service Stations, Terminals, and Centers. The tests were found to correlate highly with past experience patterns; they also had acceptable validity coefficients with training success for Center, Terminal, and Flight Service Station training; and they demonstrated ability to differentiate Air Traffic Controller from Flight Service Station job incumbents. The results of the experimental testing were analyzed at the individual item level, and these data used to create multiple forms of the test.

Since the planned initial application was a test for use as a partial basis for determining applicants' qualification for Terminal and Center job entrance above the GS-7 level, a first step was to eliminate items from consideration that had been found lacking by experts in the specialty area. The performance of virtuoso air traffic controllers was compared to the performance of entering ATC Academy students on an item by item basis. Prime criteria for item selection were (1) the extent to which each item differentiated between the two groups and (2) the difficulty level of each item. The least difficult of 100 of the most discriminating items were selected for inclusion in Form 101 - Experimental, and the next most difficult 100 items were assembled in Form 201.

In August, 1975, a battery of experimental selection tests, including Forms 101 and 201 of the ATC Occupational Knowledge Test, were administered to 109 students of the Navy Air Traffic School, Memphis. Student scores on Form 101 were distributed against their week in training when tested, and these showed increases in test performance levels as the students.

300.

Table 1

Distribution statistics and correlations
for OKT *Forms* 101 and 201, based on
109 Navy Air Traffic School, Memphis
students, 1975.

	OKT 101		OKT 201	
	Mean	SD	Mean	SD
TOTAL GROUP (N-109)	61.6	8.9	51.5	6.7
CORRELATION WITH:	r		r	
Reading	-.02		-.01	
Education Level	.27		.10	
MCAT	.42		.22	
OKT 101	---		.16	
OKT 201	.16		---	
School Grades				
Academic Average	.48		.25	
Block II Composite	.28		.16	
Block III Basic Lab	.28		.06	
Block III Advanced Lab	.27		.13	
Course Average	.43		.25	

Table 2

Distribution statistics and correlations
for OKT Form 101, based on 41 Air Force
Air Traffic School, Keesler AFB
students, 1975.

OKT 101

		Mean	Standard Deviation
Total Group (N=41)			
School Grades, Final		63.66	16.34
School Correlations			r
Block I	Basic FAA Certification		.42
Block II	Basic Operations Specialist		.15
Block III	Control Tower Operator		.32
Block IV	Radar Controller (6CA/PAR)		.68

gained experience. Grades of these students were obtained as they progressed through training. Analyses included intercorrelations and rotated factor loadings of the various tests with each other and with school grades, such as academic and laboratory grades and course average. Distribution statistics plus correlations with scores on selected variables are presented in Table 1. Mean scores on Form 201 are ten points lower than on Form 101, and the Form 201 total score correlations with other measures are similarly depressed.

In the factor analysis, student performance on the ATC Occupational Knowledge Test (OKT) showed little or no communality with students' ability to read and comprehend the material and very little relation with their mechanical-spatial aptitudes. OKT-101 had high loadings on the Air Traffic Controller Performance factor, the Control Tower Operator factor, and the Laboratory factor. OKT-201 had a similar factor structure but with lower loadings on the factors, perhaps as a function of restriction in variance, since it was a rather difficult test for this population.

A battery of experimental tests, including ATC Occupational Knowledge Test, 101 Experimental, were also administered to forty-one students at the Air Force Air Traffic Controller School, Keesler AFB, during October 1975. Distribution statistics and correlations with selected school grades are presented in Table 2. The performance of these students was quite similar to those of the Navy Air Traffic Controller School sample.

Based on the results for Forms 101 and 201 of the Occupational Knowledge Test for these groups, as well as for controllers at selected FAA Terminal and EnRoute facilities in December 1975, the best 100 items were selected to create Form 101B of the OKT. This form was part of the experimental test battery administered to FAA facility personnel and ATC Academy students in the 1977 selection study conducted by Education and Public Affairs, Inc. (Mies, Colmen, and Domenech, 1977, see Chapter 18, Part II). Distribution statistics and correlations of OKT-101B with Supervisory Assessments obtained in the 1977 study for field personnel are given in Table 3. The correlations with supervisory assessments for field ATC personnel were among the highest found for variables in the experimental test battery.

Comparable data for a group of new ATC trainees entering the FAA Academy in 1976 are provided in Table 4. As would be expected, mean scores on OKT for this trainee group were lower and the standard deviations higher than for the more experienced facility ATC specialists hired in 1969-70 or 1973-74. Correlations with the non-radar laboratory training scores (ZLab scores) for the Academy were highest for those with prior aviation experience and for minorities. Correlations for the Terminal and EnRoute options were also among the highest found for variables in the experimental test battery.

The OKT 101B was also evaluated in relation to the then current methods for granting extra credit for aviation related experience. This

Table 3

Distribution statistics and correlations
for OKT Form 101B (100 items, 97 keyed),
based on 736 FAA Facility ATC specialists,
1969-70 and 1973-74.

ATC Option and Year Hired	Correlation with Supervisory Assessment	N	Mean	SD
IFR 1969-70	.13	97	85.1	5.02
1973-74	.14	92	85.0	6.45
TOTAL	.14	189	85.0	5.75
VFR 1969-70	.13	69	82.0	9.71
1973-74	.29	100	80.6	8.39
TOTAL	.21	169	81.2	8.97
ARTCC 1969-70	.21	88	79.8	6.71
1973-74	.14	94	77.0	9.33
TOTAL	.17	182	78.4	8.26
FSS 1969-70	.01	99	75.0	9.65
1973-74	.34	97	71.5	13.04
TOTAL	.19	196	73.3	11.55
ALL OPTIONS 1969-70	.08	353	80.4	8.78
1973-74	.20	383	78.5	10.78
TOTAL	.15	736	79.4	9.91

Table 4

Distribution statistics and correlations
for OKT Form 101B, based on 803 Academy
ATC trainees, January to June, 1976.

(Academy ATC Trainees - 100 items, 97 keyed)

Academy Classes 4-8 Jan. - June 1976	Correlation with Z Lab Score	N	Mean	SD
Total Terminal and EnRoute by Subgroups		803	67.3	16.42
Aviation experience	.40	349	76.6	10.88
No aviation experience	.27	452	60.2	16.23
Men	.31	700	68.7	16.06
Women	.11	103	57.5	15.62
Minority	.36	86	67.5	16.51
Non-Minority	.29	710	68.7	16.06
Terminal	.31	372	70.9	15.44
EnRoute	.32	431	64.2	16.33

form was administered to 784 ATC trainees who entered the FAA Academy's 16-week ATC training program between July and December, 1976. All trainees completed a pre-employment questionnaire, Based on responses to the questionnaire, the trainees were assigned to one of three experience groups in accordance with evaluations of claimed experience as made by the Office of Personnel Management rating procedures. It was found that while scores on the OKT were highly correlated with claimed experience (.64), OKT had a higher correlation with successful completion of the non-radar laboratory (.25) than did claimed experience (.12). It was determined that use of an OKT score of 75 or above to assign extra credit would have resulted in a failure rate of 3.1 percent for those receiving credit, while use of the experience rating would have resulted in a failure rate of 7.7 percent for new hires who currently received extra credit for experience. The results held up for a cross-validation sample of 432 trainees who entered the Academy during 1977. The full results of this analysis of OKT have been reported elsewhere (Lewis, 1978b).

In order to obtain data on the ATC applicant groups within available testing time, a 60 item form of the OKT (101C) was developed from form 101B. This was administered to two groups of ATC applicants during September - November 1978, together with parallel forms of MCAT, after they had completed the existing (Civil Service Commission) operational ATC test battery then used by OPM (See Chapter 20). One group consisted of 5331 scheduled applicants. The second group included 669 "walk-in" applicants in testing sessions arranged to encourage women and minorities to apply for ATC work. The mean score for the scheduled applicants on the 60-item OKT (101C) was 24.9, with a standard deviation of 11.80, and for the "walk in" group, 20.1, with a standard deviation of 9.71. The OKT Form 101C was also administered to groups of ATC trainees at the ATC Academy during 1978 (See Chapter 21). From these analyses it was concluded that the OKT (Form 101B) has high reliability (.91) and that the slight loss of reliability with reduction in test length would not be of consequence.

Development of Parallel Test Forms

The purpose of the OKT is to define the domain of ATC Knowledge that is demonstrated by applicants for entry to the occupation, as a basis for granting extra credit to those who successfully pass the basic ATC selection tests administered by OPM. For operational usage, multiple forms are needed to meet retest requirements and to provide some capability for control over compromise. An essential step in the development of alternate forms was to define the types of items and numbers of items of each type to include in each succeeding form, in order to structure the domain of knowledge and make it consistent with the knowledge requirements of air traffic control. This was accomplished primarily by determining the types of knowledge that were retained after training and early apprenticeship experience and were known by nearly all full performance level controllers. Table 5 identifies the steps involved in determining the final domain of ATC occupational knowledge included in the OKT; it provides a basis for developing future comparable forms.

306.

Table 5

Item content composition of the OKT
by subject matter elements.

Form of OK Test	ATC Rules	Airport Traffic Procedures	Inflight Traffic Control Procedures	Communications Operations Procedures	Flight Assistance Service Procedures	Air Navigation & Aids to Navigation	Aviation Weather	Number of Items
1. Original FAA Basic Certification Test All Items	20	20	20	20	20	20	20	140
2. Form 16 All Items	12½	12½	25	12½	12½	13½	12½	160
3. Form 101 - Exp. All Items	20	12	28	12	2	19	6	99
4. Form 101 - Exp. Items known by 90+% of Group of Full Performance Level (N=50, 1975)	17½	9	32	12	3.5	21	5	57
5. Form 101 - Exp. Items known by 100% of Group of Full Performance Level ATC (N=16, 1970)	19	9	28	13	2	25	4	53
6. Form 101B All Items	21	12	28	12	2	19	6	100
7. New Operational Forms (Recommended)	18	10	29	13	4	21	5	100
Sub-element: 1	1	·	10	9	·	5	·	
2	·	2	3	2	·	·	2	
3	7	2	·	1	2	11	·	
4	8	1	15	·	·	·	2	
5	1	1	1	1	·	0	0	
6	·	3	·	·	2	·	1	
7	1	1						

307.

Inasmuch as the Federal Aviation Regulations require that, to become certified as a Terminal ATCS, the person must pass a written knowledge test on (1) Flight Rules in FAR Part 91 (Air Traffic Control Rules), (2) Terminal Traffic Control Procedures, (3) EnRoute Traffic Control Procedures, (4) Communications Operating Procedures, (5) Flight Assistance Service, (6) Air Navigation and Aids to Navigation, and (7) Aviation Weather, the FAA Basic Certification Test was designed to test knowledge in each of those subject matter areas. The original form of the certification test is indicated in line 1 of Table 5, which shows that an equal percentage of items (14 percent) was allocated to each subject matter area. For development of the OKT, radar questions were added to the item pool (line 2), and the best items were selected (lines 3-6) to measure candidates' possession of the knowledge required. The percentage allocation per subject matter area is presented in line 6 for version 101B of the OKT. Each of the seven subject matter areas is further defined by subject sub-elements, which are identified in Table 6. Table 7 shows the classification of items in Form 101B by element and sub-element. Numbers in the cells are item numbers. Where the item number is underlined, that item was found to be known by 90 percent or more of the 1975 sample of 50 full-performance level Air Traffic Controllers.

The subject matter areas or elements and percentage allocations per area as determined above were used to describe the types and number of items per type to be included in new alternate forms of the test. Fourteen hundred multiple choice items were drafted and reviewed, and items were selected from this pool to create eight 100-item parallel forms of the test. Twenty items were common to all forms. These forms were administered experimentally, item analysis accomplished, and the eighty best items in each form identified for keying. These eighty item forms are identified as ATC Occupational Knowledge Test 102A, 102B, 102C, 102E, 102F, 102G, and 102H, scored by a Rights Only procedure.

The forms were administered to incoming students on the first day of training at the ATC Academy beginning November 21, 1978. The mix of trainees attending the Academy varied during this period with regard to levels and types of aviation-related experience. In order to control these differences, each group of students was given two forms of the Occupational Knowledge Test. Consequently, the comparison of results for alternate forms of OKT can be made only for those tests taken by the same group of trainees.

The descriptive statistics obtained for each trainee group that took two difference forms of the tests are provided in Table 8. The means and standard deviations for the two forms taken by each group can be compared since the experience mix for that group was common for both forms taken. However, the data cannot be compared for the same form taken by different groups of trainees since the experience mix was not common.

Table 6

Categories (elements and sub-elements) of required ATC occupational knowledge.

1. AIR TRAFFIC RULES

01. Pertaining to Aircraft
02. Pertaining to Airway
03. Pertaining to Airspace
04. Pertaining to Flying Conditions (Ratings, Weather, Speed Maneuvers, Altitude, Flight Level)
05. Pertaining to Operations/Instructions
06. Pertaining to Administrative Messages (Reports, Phraseology)
07. Pertaining to Safety/Emergency

2. AIRPORT TRAFFIC PROCEDURES

01. Pertaining to Traffic Information
02. Pertaining to Traffic Patterns
03. Pertaining to Runway Operations
04. Pertaining to Flight Movement (Destination Changes, Arrivals, Departures)
05. Pertaining to Administrative Messages (Phraseology, Authorisation, Official Reports, Documents)
06. Pertaining to Meteorological Information (Ceiling, Visibility)
07. Pertaining to Safety (Messages, Warning Devices, Collisions)

3. INFLIGHT TRAFFIC CONTROL PROCEDURES

01. Pertaining to Separation (Vertical, Lateral, etc.)
02. Pertaining to Clearance (Routing)
03. Pertaining to Approach (Flight Level, Holding, Alt. Setting)
04. Pertaining to Radar (Contact, Ident, Interference, Traffic Information)
05. Pertaining to Flight Information and Data

4. COMMUNICATIONS OPERATING PROCEDURES

01. Pertaining to Flight Movement and Control Messages
02. Pertaining to Administrative Messages (Phraseology)
03. Pertaining to Meteorology Information
04. Pertaining to Clearance
05. Pertaining to Safety
06. Pertaining to Service/Maintenance

5. FLIGHT ASSISTANCE SERVICE PROCEDURES

01. Pertaining to Search and Rescue
02. Pertaining to Overdue Aircraft
03. Pertaining to Emergency Communications
04. Pertaining to Meteorological Information
05. Pertaining to Facilities
06. Pertaining to Procedures

6. AIR NAVIGATION AND AIDS TO NAVIGATION

01. Pertaining to Frequencies
02. Pertaining to Instrumentation
03. Pertaining to Position/Distance/Direction
04. Pertaining to Navigation (Time, etc.)
05. Pertaining to Navigational Aids

7. AVIATION WEATHER

01. Pertaining to Visibility/Ceiling
02. Pertaining to Weather Conditions (Temperature, Winds, Tornado, etc.)
03. Pertaining to Clouds
04. Pertaining to Administrative Messages (Phraseology)
05. Pertaining to Forecasts/Reports
06. Pertaining to Responsibilities/Procedures

Table 7

Classification of items in OKT Form 101B by elements and sub-elements. Underlined item numbers signify that 90% or more of 50 Full Performance Level Center and Terminal Air Traffic Controllers responded correctly to those items, 1975.

Sub-Element	I ATC Rules	II Airport Traffic Procedures	III Inflight Traffic Control Procedures	IV Communications Operating Procedures	V Flight Assistance Service Procedures	VI Air Navigation & Aids to Navigation	VII Aviation Weather
1	48,		1, 3, 4, 10, 15, 70, 73, 13, 74, 92	12, 18, 21, 22, 25, 11		12, 43, 44, 81, 84	
2		47, 64	4, 48	16, 16, 41, 77			29, 84, 82
3	5, 39, 53, 56, 58, 68	48, 67		13,	79,	13, 77, 18, 29, 30, 39, 54, 77, 80, 82,	
4	7, 9, 27, 41, 45, 16, 73, 76	8, 42, 43	17, 20, 22, 26, 16, 46, 90, 91, 93, 95, 94, 95, 96, 100				87, 88
5	39, 97	19	60, 98	24,		11, 31, 49, 82	
6		59, 61, 64			49,		86,
7	2, 35, 67, 63	18					
	24 items	13 items	26 items	12 items	3 items	19 items	6 items

100

The alternate forms do show some differences that require minor adjustments by shifting items of a common type but of different difficulty level form one form to another.

Reliability estimates for the various forms have been derived and are presented in Table 9. These reliabilities are considered satisfactory for operational testing use and administration.

Research with the OKT. The Lewis Study

Past practice in the selection of air traffic control trainees has been to allow extra credit for several types of prior related experience to those who passed the aptitude selection tests. The extra credit increases the applicant's standing on the register from which selection is made. Most of the credit given is for prior related air traffic control experience in the armed forces. Some credit has also been given for aircrew experience, graduate work, and superior academic achievement. An explanation of the methods of assigning extra credit for experience is included in the report by Cobb and Nelson (1974). Lewis (1978b) has reported the results of studies of the potential use of the Occupational Knowledge Test as a basis for the assignment of extra credit in the selection of air traffic controllers.

The sample in the Lewis study consisted of 1,827 EnRoute and Terminal developmental students who entered training between July 13, 1976 and February 17, 1978 (See Boone, Chapter 18, Part III). Each student passed or failed the training program and had available OKT Form 101B (100 item test) scores, the new composite experimental selection battery score, and information on ATC-related experience.

ATC-Related Experience. Information concerning experience was obtained from a biographical questionnaire administered to the developmental students on the first day of training. Based on the responses to the questionnaire, each student was given credit according to four types of experience. The four types of creditable experience were (in increasing order of relevance to ATC): a) Other experience, including air defense command, communications operator, or air traffic rating; b) Pilot experience; c) Visual flight rules (VFR) or non-radar control experience; and d) Instrument flight rules (IFR) or radar control experience. These four types of experience were used to classify each individual into one of 16 experience groups (listed below in increasing order of relevance).

        Group 1: No experience (N = 558)
        Group 2: Other (N = 34)
        Group 3: Pilot (N = 297)
        Group 4: Pilot + Other (N = 92)
        Group 5: VFR (N = 99)
        Group 6: VFR + Other (N = 27)
        Group 7: VFR + Pilot (N = 28)

Table 8

Distribution statistics for alternate
forms of the OKT.

| | (N=97) | | | (N=22) | | | |
	Mean	SD		Mean	SD		
Form A	41.95	14.05					
Form C	42.55	11.95					

| | (N=67) | | | (N=22) | |
	Mean	SD		Mean	SD
Form A	38.91	14.61		49.59	13.35
Form B	36.25	12.15		44.73	10.84

| | (N=86) | |
	Mean	SD
Form B	38.35	14.80
Form D	43.45	12.75

| | (N=384-386) | | | (N=69-70) | | | (N=22) | |
	Mean	SD		Mean	SD		Mean	SD
Form E	46.49	13.80		45.51	13.98		48.18	11.66
Form F	44.34	13.40		42.60	13.13		44.50	12.10

| | (N=407-408 | |
	Mean	SD
Form G	46.28	13.95
Form H	48.38	13.95

Table 9

Reliability of the ATC OKT

Testing Date	Student Groups	OKT 102 Taken First vs. Second		Correlation
3-6-79	Enroute N=86	E,F,G,H (80 item)	E,F,G,H (80 item)	.904
7-3-79	Enroute, Terminal, FSS, N=89	E,F,G,H (80 item)	E,F,G,H (80 item)	.893
11-21-79	Enroute, Terminal N=183	A,B,C,D (20 core items)	A,B,C,D (20 core items)	.877

313.

```
Group  8:  VFR + Pilot + Other (N = 37)
Group  9:  IFR (N = 86)
Group 10:  IFR + Other (N = 29)
Group 11:  IFR + Pilot (N = 18)
Group 12:  IFR + Pilot + Other (N = 11)
Group 13:  IFR + VFR (N - 275)
Group 14:  IFR + VFR + Other (N = 95)
Group 15:  IFR + VFR + Pilot (N = 57)
Group 16:  IFR + VFR + Other (N = 84)
```

Because the number of individuals in many of the groups was small, several of the groups were combined to form five final groups:

```
Final Group 1: No Experience or Other Experience only (N + 592)
Final Group 2: Pilot or Pilot + Other (N = 389)
Final Group 3: VFR and any other experience except IFR (N = 191)
Final Group 4: IFR and any other experience except VFR (N = 144)
Final Group 5: IFR + VFR and any other experience (N = 511)
```

Composite Experimental Selection Battery Score. The new experimental battery composite score was based on the weighted sum of three tests: CSC-Arithmetic Reasoning, CSC-Abstract Reasoning, and the Multiplex Controller Aptitude test. The weighted test composite was divided by three to form a new selection score with a range of 0 to 100, which approximated the transmutation anticipated when the new weighted score would be used operationally by the Office of Personnel Management.

Measures of Training Success. Training success was measured by pass-fail in the Academy and by a computed measure called ZLab (See Chapter 18, Part III). The students who successfully completes the EnRoute or Terminal ATC training course with phase composite scores of 70 or better were coded as passes. Those who did not complete training because of composites of less than 70 were coded as failures. Those who did not complete training because of composites of less than 70 were coded as failures. Those who withdrew from training for any reason were eliminated from the sample.

ZLab is the numeric grade from the non-radar laboratory portion of Academy training, converted to standard score form. This score was standardized to minimize differences in laboratory grading procedures between the two different ATC options. The data from the two ATC options were combined and ZLab was used by Boone as the criterion variable for the development of the new selection score. Although ZLab is a training performance measure it has demonstrated a significant relationship to performance in the field as measured by field supervisors' ratings of the potential of Academy graduates who had been in the field for at least six months. The correlation between supervisor rating and ZLab was .227 ($p < .001$) for a sample of 1,689 EnRoute and Terminal Academy graduates.

314.

Results. As can be seen from Table 10, the overall OKT (100 item test) mean was 66.96 and the means by experience groups ranged from a low of 51.49 for the Other experience group, to a high of 78.0 for the VFR and IFR experience group. The co-relation of OKT with experience was high (.61), and OKT had a higher correlation with ZLab than did rated experience (.22 versus .11). This indicates that while OKT was closely related to experience, it was even more predictive of success than was a score based on rating of experience.

Table 10 also shows the distribution and failure rate in training for six OKT score ranges for the five experience groups and for the total sample. As can be seen in this table, the first experience group had the highest failure rate (26.9%) and the VFR only and the VFR plus IFR experience groups had the lowest failure rates among the experience groups (13.2% and 14.3%). If OKT scores had been used to assign extra credit, and a minimum score of 70 on the OKT were required, 957 students would have received extra credit for experience and only 11.5% of these would have failed. By contrast, based on the experience categories in Table 10, 1,235 student (all categories except Other) would have received extra credit and 15.5% would have failed.

As shown in Table 11, if experience credit were assigned to individuals who scored 70 or above on the OKT, and to individuals in all but the first experience category, 511 individuals would have received no extra credit on either the OKT or experience category (and their failure rate would have been 28.8%), while 876 individuals would have received extra credit on both OKT and experience (and their failure rate would have been 11.2%). Thus, using either OKT or experience would have resulted in agreement to allocate extra credit to 1,387 individuals, comprising 75.9% of the sample. The two methods of assigning extra credit would have disagreed on 440 students in the sample of these, the OKT would have assigned extra credit to 81 individuals who would not have received credit for experience (and who would have had a failure rate of 14.8%), and the experience crediting system would have assigned credit to 359 individuals who would not have received OKT credit (and their failure rate would have been 26.2%). These data indicate that the OKT is an effective means for the assignment of extra credit for ATC-related experience. In fact, by using the OKT with a minimum score of 70, in place of ratings based on experience categories, the failure rate for those receiving credit would be reduced from 15.5% to 11.5%.

Implementation of the OKT

Prior to the adoption of the new selection test battery in 1981, the procedure for allocating extra credit for experience was to assign either 0, 5, 10, or 15 points for the various experience categories. Based on the evidence cited above, it was concluded that the knowledge retained from ATC-related experience could be measured by OKT scores and that extra credits should be assigned to individuals who scored 70 or better on the OKT. The failure rate for developmental students who scored 70 or higher

315.

Table 10

Failure Rates by OKT Score Ranges, and OKT Means by Experience Groups

Score	Other Total	Other Pct. Fail	Pilot Total	Pilot Pct. Fail	VFR Total	VFR Pct. Fail	IFR Total	IFR Pct. Fail	VFR + IFR Total	VFR + IFR Pct. Fail	TOTAL Total	TOTAL Pct. Fail
0-59	430	29.3	67	26.9	25	32.0%	11	27.3%	23	43.5%	556	29.7%
60-64	37	27.0	47	23.4	19	42.1%	9	33.3%	23	34.8%	135	29.6%
65-69	44	25.0	55	9.1	21	28.6%	13	23.1%	46	23.9%	179	20.1%
70-74	34	20.6	77	13.0	36	19.4%	18	16.7%	61	16.4%	226	16.4%
75-79	22	9.1	67	10.4	37	13.5%	34	11.8%	117	14.5%	277	12.6%
80+	25	12.0	76	10.5	53	13.2%	59	5.1%	241	7.1%	454	8.4%
TOTAL	592	26.9	389	15.2	191	21.5%	144	13.2%	511	14.3%	1827	19.2%
Means on OKT	51.49		69.94		72.25		75.28		78.0		66.96	
std. dev.	14.8		10.9		11.4		9.5		9.7		16.3	

316.

Table 11

Distribution of OKT credit by experience credit and fail rate.

	No OKT Credit	OKT Credit	Total
No experience credit	N = 511 28.8% fail	N = 81 14.8% fail	N = 592 26.9% fail
Experience credit	N = 359 26.2% fail	N = 876 11.2% fail	N = 1253 15.5% fail
Total	N = 870 27.7% fail	N = 957 11.5% fail	N = 1827 19.2% fail

317.

was 11.2% while the failure rate for those who scored below 70 was 27.2%. As demonstrated in Table 10, the higher the OKT score, the smaller the proportion that failed the training program. As a result, instead of basing the award of extra credit on experience group, it seemed appropriate to add additional points in relation to the magnitude of OKT scores. The values adopted, by score range, are described in Table 12. This method of extra point allocation is similar to the experience group system, and as with that system, the extra points were to be added to the new experimental battery score.

Thus far the data reported have simply compared prediction based on the OKT with prediction based on rated experience. However, in order to demonstrate that the OKT is a truly useful selection instrument, it was necessary to investigate whether it contributed to prediction over and above the predictions made by the experimental battery.

Boone (1979a, Chapter 18, Part III) had shown that the multiple regression coefficient for the experimental battery (not including the OKT) was .54, accounting for 29% of the variance in ZLab. After correcting the correlations of OKT with ZLab and the other experimental battery test scores for restriction of range, using Thorndike's (1949) formula 7, it was appropriate to estimate the magnitude of the multiple regression coefficient with the OKT included in the analysis. The new multiple regression coefficient, calculated by including OKT along with the three tests already in the battery, was .61 and accounted for 36.7% of the variance in ZLab. This increase is significant (F=204.349, p<.001) and indicates that the addition of extra points using OKT would significantly improve prediction. It supports the recommendation that OKT should be used to add extra credit for ATC-related experience.

The effect of adding extra credit based on the OKT, by the system described in Table 12, on the distribution of the selection scores for the new battery is shown in Table 13. It can be seen that not only is there a clear linear relationship between the selection scores and failure rates, but also that the range of failure rates and the consistency of the dropoff are greater when OKT credit is counted. In practice, most applicants selected off the register have had selection scores of 90 or higher. If no extra credit were added for OKT scores, as shown in Table 13, 1,002 developmental students would have had scores of 90 or greater, and the overall failure rate for these 1,002 individuals would have been 12.7%. However, when extra credit for OKT was added, these same 1,002 individuals still had scores of 90 or higher, and the scores of an additional 292 individuals were raised so that they could be included in the 90+ score range. The overall failure rate remained around 13% for the enlarged group with scores greater than 90, but the most dramatic change occurred in the very highest group. Without extra credit for OKT, 425 individuals would have had scores of 100 and 11.1% would have failed; with OKT extra credit, an additional 339 individuals would have been included in this category and the failure rate would have been reduced to

318.

Table 12

Values adopted by score range
OKT 100 item test

OKT score range	Percent failing	Points allowed
0 - 69	27.7%	0
70 - 74	16.4%	5
75 - 79	12.6%	10
80+	8.4%	15

Table 13

Failure rates by score ranges on new battery
with and without extra credit on OKT.

New Selection Battery Score

Score	With No credit		With OKT credit	
	N	% fail	N	% fail
0-69	67	47,8%	40	55.0%
70-74	81	35.8%	49	40.8%
75-79	155	28.4%	87	39.1%
80-84	242	27.3%	159	34.6%
85-89	280	18.9%	198	26.3%
90-94	285	14.7%	256	21.9%
95-99	292	13.0%	274	16.4%
100 +	425	11.1%	764	8.8%
TOTAL	1827	19.2%	1827	19.2%

320.

8.8%. This indicates that adding extra credit of either 5, 10, or 15 points for OKT scores is a valuable way to increase the proportion of successful individuals in the upper ranges of the CSC register.

SUMMARY

The research reported has shown that the OKT is highly correlated with ATC-related experience, and does a better job than rated experience in predicting FAA Academy success. By using the OKT in place of experience to assign extra credit for ATC-related experience/knowledge, the Academy attrition rate for those receiving credit would be decreased from 15.5% to 11.5%. The major difference realized by the allocation of extra credit based on the OKT instead of experience is that 359 subjects, with an attrition rate of 26.2%, would have been denied extra credit that they would have received under the system based on experience. In addition to being superior to rated experience in the prediction of academy training success, the OKT also contributes significantly to prediction when added to the new selection score. It increased the variance accounted by the new selection score by a significant amount, from 29.5% to 36.7% (p<.001).

The OKT appears to be a measure, not merely of the amount and type of an applicant's experience, but also of the quality of that experience, in terms of useful knowledge retained. It enables the identification of those individuals who have ATC-related experience on paper, but who do not have the ATC-related knowledge that should reflect that experience. The use of the OKT to replace the assessment of previous experience is an extremely cost-effective means for the assignment of extra credit. The OKT can be machine scored, whereas the rating of experience was not; machine scoring not only reduces the probability of mathematical error in the addition of extra credit, but it eliminates the time-consuming review of CSC Form 171's previously used to determine the applicants' relevant work experience. The result is a considerable reduction of work hours required to assign extra credit as well as an increase in both the mathematical accuracy of the calculated scores and the predictive accuracy of the modified selection scores. Based on the results of the Lewis studies, it was strongly recommended that the OKT be used to assign extra credit, using the system described in Table 12.

321.

REFERENCE NOTES

1. Boone, J. O. Proposed implementation of a new selection battery for the selection of air traffic controllers. Unpublished FAA document, 1979.

2. Cureton, Louise Witmer. Early identification of behavior problems. Final Report, American Institutes for Research, January 1970.

Chapter 17

PERSONALITY ASSESSMENT OF ATC APPLICANTS

John J. Convey

Since World War I, tests have been a favored tool for personnel se-
lection. Shortly after the War, personality inventories began to be
used in some employment settings. The 1930s and 1940s saw a dramatic
increase in the use of personality inventories in employment. However,
by 1950, many psychologists had questioned the use of personality in-
ventories, and these had come under open attack by the mid-1950s. In
fact, one of the principal motivations for the 1954 Technical Recom-
mendations for Psychological Tests and Diagnostic Techniques, by the
American Psychological Association, was the overuse of personality
testing for employment selection (Novick, 1982). By the 1970s, the use
of personality tests had waned considerably, most probably due to the
intense attack to which they were subjected as a result of their low
validity in predicting job performance and their apparent invasion of
privacy.

A basic rationale for choosing tests for worker selection is based
on the assumption that objective measurement is superior to subjective
decision-making. This assumption has been debated for years and has
been the object of a sizeable literature (see Goldberg, 1970; Meehl,
1954). However, most writers on testing would agree that human judg-
ment plays a role even when test scores are the primary source of in-
formation about applicants (Wigdor & Garner, 1982). This would seem to
be particularly true with regard to personality inventories and their
use by trained psychiatrists and psychologists. Despite the fact that
personality inventories have traditionally had low validities in the
prediction of job success for workers, these tests have been used exten-
sively by clinicians in screening individuals for various forms of psy-
chopathology.

Personality assessment has been used formally by the Federal Avia-
tion Administration (FAA) in conjunction with the Air Traffic Controller
(ATC) program since 1965. After a brief period of experimental use with
employed controllers, the Sixteen Personality Factor Questionnaire (16PF)
was included as part of the medical screening program for all ATC appli-
cants. From the outset, the 16PF was intended to be used as a case
finder; that is, an indicator of emotional instability for referral of
screened applicants for psychiatric and psychological evaluation. Since
May of 1975, a score based on a 38-item subset from the two 187-item
forms of the 16PF has been used as the case finder. This 38-item sub-
set is primarily a measure of anxiety. The purpose of this chapter is
to document the use of personality assessment, particularly the 16PF,
within the ATC program since 1965. Of particular interest is the docu-
mentation of the development of the new scoring procedure and its use.
In addition, the potential of other sources of personality information,

particularly life history data, which could be useful in the selection of ATC applicants is examined.

The chapter consists of nine additional sections: Introduction to Personality Assessment; Structured Self-Report Personality Inventories; The 16PF; The Use of the 16PF with ATCSs since 1965; The Personality Profile of the ATCS; The Validity of the 16PF for the ATCS; Life History Data; Practices in Other Countries; and Summary.

INTRODUCTION TO PERSONALITY ASSESSMENT

Sundberg (1977) defined personality as ". . . the system whereby the individual characteristically organizes and processes biophysical and environmental inputs to produce behaviors in interaction with the larger surrounding systems" (P. 12). According to Sundberg, three major purposes of personality assessment involve image-making, decision-making, and theory-building. In image-making, an attempt is made to develop descriptions and images of the person. This is often done through the construction of personality profiles. In decision-making, the purpose is to make decisions concerning the relationship of the person to his or her environment. The decisions may involve either personnel or clinical situations, and are basically either for selection or classification. In theory-building, hypotheses about personality are tested and theory is related to practice.

The primary rationale for the use of personality assessment within the FAA for the Air Traffic Control Specialist (ATCS) to date has been for decision-making. The fundamental questions appear to be: Which applicants will not be able to function properly as an ATCS because of severe personality problems, particularly an abnormally high level of anxiety? What personality information is useful for the effective screening of ATCS applicants? How valid is the current screening procedure in the identification of individuals in need of further assessment? In an attempt to understand the personality of the typical ATCS, much of the previous research using ATC data has involved the image-making purpose of Sundberg. Researchers have explored questions such as: What type of personality does the typical ATCS possess? How does the personality profile of an effective ATCS differ from the profiles of other occupational groups and professions? How does the personality of a typical ATCS relate to the performance of his or her job? Recently, Hopkin (1980) called for research with ATCSs which would satisfy Sundberg's third purpose, theory-building. Hopkin indicated that researchers ought to start from basic psychological concepts and derive hypotheses from them for testing in an air traffic control context. While Hopkin suggested that these concepts could come from cognitive psychology, it appears that some could also come from personality theory.

The most widely used methods of personality assessment are self-report inventories, projective tests, and clinical interviews. Other methods of assessment include physiological techniques (biofeedback, polygraphs, telemetry, and voice stress analysis), situational simulations, biographical inventories or weighted application blanks, selection interviews, and

background and reference checks. These methods are normally discussed in detail in text books on personality assessment (for example, Sundberg, 1977; see also, Frank, Lindley & Cohen, 1981). Since this chapter deals with the personality assessment of ATC applicants, only those methods that are presently used with applicants or which would be feasible to use are discussed.

Presently, the 16PF, a self-report personality inventory, is routinely given to all ATCS applicants. From a feasibility standpoint, information from biographical inventories and background and reference checks could also be used, although such information is not currently used in a formal way. Clinical interviews are also used, but only for a very small percentage of applicants not cleared by the 16PF. Even though projective tests continue to be widely used in clinical settings, they are not feasible first-stage assessment methods for ATCS applicants, in part because of the difficulties involved in scoring and interpreting such instruments and in part because of their lack of demonstrated validity. There has been a decline in the popularity of projective techniques in the last twenty years, at least partly due to a shift from psychoanalytic theory to a behavioral personality theory (Sundberg, 1977). The remaining methods, physiological techniques, simulations, and selection interviews, are not feasible for ATCS screening.

STRUCTURED SELF-REPORT PERSONALITY INVENTORIES

Background

Among the many structured self-report personality measures, those dealing with psychopathology and adjustment are most relevant for this chapter. Other types of structured personality inventories deal with vocational interests, self-concept, locus of control, personal value, masculinity-femininity, and the like. This section draws heavily on the historical overview of personality scales and inventories by Goldberg (1971).

The forerunner of all personality and adjustment scales was Woodworth's Personal Data Sheet, developed in 1917. Woodworth's major contribution was the production of a single scale by combining the responses to questionnaire items into one score (Goldberg, 1971). The resulting scale, called Psychoneurotic Tendencies, was developed to screeen emotionally unstable soldiers during World War I. Items from Woodworth's scale were adopted over the years by many psychologists, including L. L. Thurstone and T. G. Thurstone for their Personality Schedule, which appeared in 1930. This scale formed the basis for the Neurotic Tendencies in the very popular Bernreuter Personality Inventory which was used by some companies in the late 1930s to identify stable, loyal, and productive workers (Hale, 1982).

Prior to 1920, items for personality inventories were assembled using an intuitive approach (Goldberg, 1971). Basically, this involved clustering

325.

the items into sets to relate to a particular construct. Sundberg (1977) referred to this developmental approach as a judgmental or rational approach, while Anastasi (1968) termed it a content validation approach. By the mid-1920s, the intuitive approach was being combined with what might be called an internal approach. This involved retention of intuitively developed items in scales, based on the correlation of the item with the scale.

A different approach to item development and selection was used by Pressey in 1921 to assess scholastic potential and by Strong in 1927 in the development of the Vocational Interest Blank. This approach has been called external by Goldberg (1971), empirical by Anastasi (1968), and group contrast by Sundberg (1977). An item was retained in an inventory if it discriminated adequately between existing criterion groups. The first use of this approach to develop adjustment scales was made by Humm and Wadsworth in 1935, in their Temperament Schedule. This inventory yielded scores for seven components of adjustment versus psychopathology. The scales were based on item analysis, using criterion groups of patients and normals who had been judged as being extreme on the respective components. This same strategy was used by Hathaway and McKinley in the development of the Minnesota Multiphasic Personality Inventory (MMPI) published in 1943, and by Gough for the California Psychological Inventory (CPI) published in 1956.

A new type of internal approach emerged in the mid-1930s when Guilford began a series of factor-analytic investigations of personality scales. These investigations led first to the Nebraska Personality Inventory published in 1934, and ultimately to the Guilford and Zimmerman Temperament Survey published in 1949. During this same period, Cattell was undertaking his factor-analytic investigation of the "total personality sphere", one outcome of which was the publication in 1949 of his 16PF. Both Guilford and Cattell developed their scales on normal subjects; and both conducted subsequent research that included tests of whether their scales differentiated psychiatric patients from other groups. A third researcher closely identified with factor-analytic personality scales was H. J. Eysenck. Unlike Guilford and Cattell, Eysenck had been constructing factor scales aimed from the start at differentiating clinical groups from other groups. In 1959, Eysenck published the Maudsley Personality Inventory (MPI); the Eysenck and Eysenck Personality Inventory (EPI), in 1966, was a modified version of the MPI.

The internal approach to scale construction, particularly the use of factor analysis, has dominated other approaches since the end of World War II. A notable exception has been the CPI, which was based on the MMPI. An apparent advantage of an internal approach is that the scales produced can be made to be relatively independent of each other (Cattell's 16PF is an exception), while inventories constructed using an external approach are not usually sensitive to this and tend to have highly correlated scales.

326.

Comparative evaluation of these approaches is difficult. Theoretically, instruments based on an external approach might be expected to have higher criterion-related or predictive validity than those based on an internal approach, but Goldberg (1972) has shown that scales produced by each method were equally effective in predicting an external criterion. Loevinger (1972) considered the use of internal approaches as evidence for one aspect of construct validity; nevertheless, she did admit that ". . . it is fallacious to assume that any factorial method automatically supplies test scores or trait concepts that are robust under demographic variation" (P.56). At present, the 16PF, which is based on an internal approach, and the MMPI and the CPI which are based on an external approach, are widely used in clinical settings, and are the three most widely used measures of adjustment in non-clinical settings (Frank et al., 1981).

Technical Properties

The major advantages of structured self-report personality inventories are that they are standardized and yield scores which can be compared to some norm group or groups. These instruments are easy to administer, are adaptable to objective scoring, and yield scores which can easily be verified. Furthermore, they are likely to detect subjects who have some degree of psychopathology (Buchanan, David & Dunnette, 1981). In addition, most inventories have scales to detect subjects who have attempted to falsify information or to present themselves differently from what they are.

The major disadvantages of these inventories are ambiguity of items, and acquiescence and faking on the part of the respondents (Cronbach, 1970). Various interpretations that can be applied to wording, especially frequency wording, and the need to deal with the hypothetical "typical" situation rather than specific, well-defined situations contribute to the ambiguity of the items. Furthermore, many subjects tend to pick the socially desirable response, regardless of whether that is an appropriate response for them. This is a particular problem when using such instruments in employment situations, especially at the applicant level. Finally, some subjects deliberately portray themselves in an unfavorable manner. This is less of a disadvantage in employment situations, since such action may be indicative of some personality disorder.

Additional considerations need to be taken into account when inventories of this nature are used in employment situations. Butcher (1979) indicated that some items may be perceived as an invasion of privacy or as not being relevant to the particular job. Butcher warned: "Situations in which the individual takes a personality test for someone else's benefit (employment selection, research, etc.) require special precautions and more carefully planned insturctions and debriefings" (Butcher, 1972, p. 11). Furthermore, on the basis of present knowledge, the use of such inventories as the sole basis for selection for employment is not warranted. Cronbach (1970) indicated that personality inventories tend to be poor predictors of employee performance, and no empirical data have become

available in the last 12 years to change that assessment. It should be noted, however, that even correlations between scores on ability tests and occupation performance criterion measures are typically quite modest, commonly in the .20s (Linn, 1982).

THE 16PF

Description

The 16PF (Cattell, Stice, & Eber, 1949) in an inventory, developed using factor-analytic methods, designed to assess the personality traits of normal subjects. This inventory has six forms (A, B, C, D, E, and F) and two revisions. Forms A and B have 187 items each and assume a high school reading level. Forms C and D are shorter (106 items each), somewhat less reliable than Forms A and B, and assume a pre-high school reading level. Forms E and F are for low literate examinees.

The inventory contains 16 subscales, representing 16 primary factors. Forms A and B each contain from 10 to 13 items for each of the 16 subscales. A description of the primary factors is given in Table 1. For each item, the examinee is give a choice of three responses. Only items on scale B (intelligence) have a correct answer; the remaining items do not have "right" or "wrong" answers as such. Normally, an examinee takes both Form A and Form B. Scores for the factors are scaled using a sten transformation (mean of 5.5 and standard deviation of 2.0), and are plotted as a profile. Faking and acquiescence are partially controlled by using items in which the socially desirable response is not obvious. The inventory has a Motivational Distortion (MD) scale which can be used to identify extreme attempts at falsifying the responses.

Based on the correlations among the 16 primary factors, Cattell derived four major and several minor second-order factors for which scores are provided. The major second-order factors are Anxiety, Introversion-Extraversion, Poise, and Independence. The second-order factors resulted from factor analysis of the 16 by 16 primary factor correlation matrix obtained from an oblique factor solution to the original item set, using the data from the standardization sample. The second-order factors are more general than the primary factors and consist of linear combinations of the primary factors. Two of the second-order factors, Anxiety and Introversion-Extraversion, are generally recognized as being stable factors of the 16PF (Bolton, 1977). These traits are addressed in most personality inventories designed to measure psychopathology and adjustment.

Background

Cattell's research on personality (1946, 1950, 1957) began with the conceptualization of the "total personality sphere". Cattell assembled a comprehensive list of personality traits from words descriptive of personality aspects compiled by Allport and Odbert (1936) and from reviewing the psychiatric and psychological literature. Approximately 18,000 trait-descriptive words were reduced to 171 by eliminating duplication and

Table 1
Personality Description of the High and Low
Scorers on the 16 P.F. Factors

Factor	A Person with a Low Score is:	A Person with a High Score is:
A	Reserved, detached, critical, cool	Outgoing, warmhearted, easy-going, participating
B	Less Intelligent, concrete-thinking	More Intelligent, abstract-thinking, bright
C	Affected by Feelings, emotionally less stable, easily upset	Emotionally Stable, faces reality, calm
E	Humble, mild, obedient, conforming	Assertive, independent, agressive, stubborn
F	Sober, prudent, serious, taciturn	Happy-Go-Lucky, heedless, gay, enthusiastic
G	Expedient, a law to himself, by-passes obligations	Conscientious, preserving, staid, rule-bound
H	Shy, restrained, diffident, timid	Venturesome, socially bold, uninhibited, spontaneous
I	Tough-Minded, self-reliant realistic, no nonsense	Tender-Minded, dependent, overprotected, sensitive
L	Trusting, adaptable, free of jealousy, easy to get on with	Suspicious, self-opinionated, hard to fool
M	Practical, careful, conventional regulated by external realities, proper	Imaginative, wrapped up in inner urgencies, careless of practical matters, bohemian
N	Forthright, natural, artless, sentimental	Shrewd, calculating, worldly, penetrating
O	Placid, self-assured, confident, serene	Apprehensive, worrying, depressive, troubled
Q_1	Conservative, respecting established ideas, tolerant of traditional difficulties	Experimenting, critical, liberal, analytical, free-thinking
Q_2	Group-Dependent, a "joiner" and good follower	Self-Sufficient, prefers own decisions, resourceful
Q_3	Casual, careless of protocol, untidy, follows own urges	Controlled, social' recise, self-disciplined, compulsive
Q_4	Relaxed, tranquil, torpid, unfrustrated	Tense, driven, overwrought, fretful

NOTE: These descriptions are found in About the 16 P.F., published by the Institute for Personality and Ability Testing, Champaign, Ill., a four-page, uncopyrighted brochure.

ambiguity. These 171 characteristics were used to rate a sample of persons and the resulting ratings were factor analyzed by Cattell. At first, 12 factors were identified in the rating data; later, these were represented in questionnaire items along with new items, and the four Q scales were added in subsequent research.

Since the appearance of the 16PF in 1949, there has been considerable controversy concerning the inventory. Bolton (1977), in an overview of this controversy, indicated that criticism of the 16PF has progressed from sporadic attacks in the 1960s to full-scale assaults in the 1970s. The controversy centered around the existence of the factors as defined by Cattell. Investigations in which the 16PF was used have either supported Cattell almost unequivocally or have found little value in his work. Bolton (1977) indicated that most of the former group have tended to be Cattell's students and colleagues.

Criticism of the 16PF began with Levonian's (1961) questioning of the psychometric adequacy of the instrument. Levonian noted that the scale correlated with more items outside their respective scales than within, and concluded that the scales of the 16PF lacked homogeneity. Sells, Demaree, and Will (1968, 1970, 1971) conducted a combined factor analysis of 600 items from the Guilford and Cattell inventories. Their results confirmed the heterogeneity of the Cattell (and also Guilford) factors. Eysenck, White and Soueif (1969) factor analyzed 99 16PF items selected by Cattell and concluded that". . . Cattell's questionnaires. . . should not be used to measure the Cattell primary factors, whose existence received no support from this investigation" (p. 228).

Howarth (Howarth & Browne, 1971; Howarth, 1976) added further evidence on the same theme. Howarth and Browne performed an item factor analysis of the 16PF and produced a ten factor rotated solution. They concluded that the 16PF does not measure the primary factors that it purports to measure and that ". . . Cattell's questionnaire factor system has been developed on the basis of inadequate investigation of the primary factors" (P. 138). Eysenck (1971, 1972) seized upon the Howarth and Browne results and used them as a basis for denunciation of the 16PF. Eysenck (1972), who is an advocate of two super-factors, E, Extraversion-introversion, and N, Neuroticism, denied the existence of Cattell's primary factors and argued only for the existence of second-order factors, which he claimed were really his E and N. Howarth (1976) reanalyzed Cattell's original data and extracted six factors instead of the earlier 12 factors. These were labeled cooperativeness, surgency, emotional maturity, extraversion, superego, and adjustment. Howarth claimed that these factors agreed with work carried out independently of Cattell's laboratory.

Cattell's (1972, 1973) response to his critics was that they failed to use factor analysis properly. He claimed that other researchers consistently underfactored the item matrix and, as a result, did not replicate his structure, but often produce "pseudo" second-order factors. Some of Cattell's students (DeYoung, 1972; Vaughn, 1973) criticized the Howarth and Browne (1971) study for methodological inadequacies.

On the proponent side, Cattell, Eber, and Delhees (1968) found support for 12 of the 16 factors (E, M, N, and O were not clearly defined). Karson and O'Dell (1974a), using data for air traffic controllers, found only weak support for Cattell's framework, but concluded that their results were not as discouraging as some investigators have suggested. More recently, Turner and Horn (1977) found results for 489 adults that were consistent with Cattell's. Golden (1978) found that the second-order factor structure for Japanese subjects differed from Cattell's structure, but that the structure for those of European ancestry did not. Bolton's (1977) assessment was that, for the most part, good support is available for Cattell's second-order factors and most of his primary factors, and he concluded that the 16PF compared favorably to other personality inventories.

Perhaps the evaluation of Cattell's work by Hall and Lindzey (1970) summarizes the controversy the best:

> Great chunks of Cattell's theoretical structure rest
> on shaky empirical findings. But great chunks of any
> comprehensive personality theory must rest on shaky
> empirical underpinnings. Good personality research is
> slow and expensive. It is to Cattell's immense credit
> that so much of his theory has as much empirical ground-
> ing as it does. It must be conceeded to Cattell's dis-
> credit, however, that he frequently claims a stabler
> empirical foundation than actually exists for his con-
> structs. Cattell the tactician here does Cattell the
> strategist a considerable disservice. (pp. 408-409)

THE USE OF THE 16PF WITH ATCSs SINCE 1965

The use of the 16PF within the ATC program can be viewed in three, almost equal-length phases. The first phase, which lasted from 1965 to approximately 1970, consisted of the initial administration of the 16PF to working controllers and subsequently to applicants; this resulted in some preliminary explication of its value for producing meaningful personality profiles and for its use as a screening instrument. The second phase lasted from 1971 to approximately 1975 and was characterized by two major developments. The first involved criticism of the internal structure of the 16PF and questioning of the utility of the profile produced. This subsequently led to the second development, which involved redefinition of the second-order anxiety factor and the development of a brief, efficient scoring procedure to measure that factor. The third phase, from 1975 to the present, consists of the use of the revised scoring procedure for routine screening, in the case-finder sense, of ATC applicants. The three phases are discussed in this section.

Phase I

On October 15, 1965, in an executive order issued by D. B. Thomas, the acting administrator of the then Federal Aviation Agency, the use of

a psychological test battery was prescribed as part of the medical examination under the new health program for ATCSs. The order applied only to applicants for, and holders of, ATCS positions in agency field facilities, and did not affect personnel in supervisory and administrative positions who did not handle air traffic directly. Prior to this date, ATCSs who worked in terminals were required only to meet second-class airmen medical standards on an annual basis. Other controllers employed by the agency had been required, as a condition of initial employment, to possess valid second-class airmen medical certificates, but were not required to undergo periodic examinations. The psychological test battery, as part of the medical examination, was intended to screen for personality disorders that might detract significantly from an individual controller's ability to function according to performance standards.

Prior to 1965, it had been generally recognized that few occupations on a day to day, minute to minute, second to second basis, were as demanding as that of an air traffic controller. In 1956 and 1957, The Flight Safety Foundation, Inc., a non-governmental organization, recommended that standards of medical fitness for ATCSs be tailored to the demands of the job. This recommendation coincided with the opinion of the FAA medical staff (Siegel, 1966). Appropriated funds required to implement a comprehensive medical examination of ATCSs were made available in late 1963. A set of medical standards, based on past research studies on ATC populations and on an analysis of the work demands in control facilities, was officially adopted by the U. S. Civil Service Commission in April, 1965.

After reviewing the available psychological test instruments, particularly with regard to validity and ease of administration (Siegel, 1966), the 16PF was selected as the psychological battery component of the medical examination. From the outset, the test was intended to be used as a case-finder (similar to the use of laboratory tests such as the electrocardiogram or chest x-ray) which could be used to identify individuals thought to require comprehensive psychiatric and psychological assessment. Because the 16PF was to be used in a decision framework, the establishment of cut-off scores was necessary. Based on the use of the 16PF with other poulations, the score on the second-order anxiety factor was selected as the determiner of a more comprehensive assessment. A cut-off score of 9.5 (on a scale of 10) was established, with the expectation that not more than one percent of ATCSs would exceed it. Reighard (Note 1) indicated that prior clinical experience suggested that the use of this score would identify most problem cases while avoiding the identification of too many false positives (i.e., individuals without identifiable psychiatric or personality problems who scored at or above 9.5). It was intended that any controller whose score reached or exceeded 9.5 on the anxiety scale would be interviewed by a flight surgeon, usually in the region that employed, or intended to employ, the controller, and referred for a complete psychiatric and psychological evaluation.

In early 1966, the 16PF was administered to over 6200 ATCSs in center facilities. Later in the same year, over 4400 ATCSs from towers were

332.

tested. By January, 1967, the 16PF had been administered to over 12,000
ATCSs and applicants (Note 2). Reighard (Note 1) reported that 151 con-
trollers from this first round of testing scored at or above the cut-off
score of 9.5 on the second-order anxiety factor, and thus were identified
for further assessment. This number represented 1.2% of the controllers
examined and was approximately the number expected. Of these 151 control-
lers, 60 (39.7%) were cleared by regional flight surgeons without referral
for more formal psychiatric assessment. (It should be noted that the vast
majority of the over 12,000 tested were already employed controllers.) Of
those controllers who were referred for further assessment, 15 were found
to have a severe disturbance and were recommended for termination, while
16 had disturbances not severe enough to require permanent removal from
duty. Arrangements were made for the latter group to receive appropriate
treatment and followup, either while temporarily removed from duty or while
continuing to serve as controllers.

The introduction of the 16PF as part of the medical screening program
for the ATCSs came at a time when there was widespread controversy concern-
ing the use of personality tests in employment situations, in general, and
for federal employment, in particular. In 1964, congressional hearings
were held on the use of personality inventories by the government. In
1965, John W. Macy, Jr., the U. S. Civil Service Commissioner, testified
to Congress that the use of present personality tests and inventories was
not justified as a selection method (Hale, 1982). William McKee, the FAA
Administrator, presented some preliminary evidence for the validity of the
16PF in a memo to Macy (Note 2). In addition to presenting data similar
to that given by Reighard (Note 1), McKee cited three cases of breakdown
occurring before a psychological evaluation had been completed. In each
of these cases, the individual was placed in the mandatory followup cate-
gory as a result of the scores on the 16PF. While the evidence presented
by McKee was based on a very small number of subjects, the fact remains
that the 16PF did identify some individuals with severe personality distur-
bances who were already functioning as air traffic controllers and who had
not been previously identified as having such problems.

In the same memo, McKee also addressed the concern of the U. S. Civil
Service Commission for the manner in which the test was administered, the
confidentiality of the data, and potential biases in the items. McKee
gave assurance that the administration of the instrument ensured the con-
fidentiality of the individual answer sheets, and that individuals were
informed that they could omit any questions they found personally objec-
tionable. In addition, McKee informed Macy that he had arranged for 25
items of the 16PF to be revised by the Institute for Personality and Abili-
ty Testing, which published the 16PF, because of objectionable content.
The sanitized version of the 16PF was available by mid-1967 for use with
all new ATC applicants.

The testing of ATC applicants was not interrupted, but continuation
of the screening on an annual basis for working controllers was postponed
after the first round of testing, mainly because of ". . . some difficulties

encountered during the first round of testing" (Reighard, Note 1).
Another reason contributing to the postponement was the resignation of
the clinical psychologist who was the program manager for the personality
assessment portion of the medical program. Since the initial testing in
1966, there has been no periodic personality assessment of the employed
controllers. The only use of personality assessment has been as part of
health related studies, such as that by Rose, Jenkins, and Hurst (1978)
and those on stress conducted by the Civil Aeromedical Institute (Smith,
1980). Another reason for not continuing the testing of employed con-
trollers was undoubtedly the controversy concerning the use of personality
testing for federal employees. The use of the 16PF with applicants con-
tinued, however, on a case-finder basis.

Phase II

Phase II is characterized by the contributions of John Dailey re-
garding the use of the 16PF for ATC applicants. Dailey's awareness of
the literature and his own analysis led him to question the validity of
the primary structure of the 16PF. He was convinced that the 16PF did
measure at least the anxiety factor quite well, but he questioned the
credibility of the results of the 16PF, especially for the anxiety factor,
when the instrument was taken as a condition for employment. Dailey ana-
lyzed 16PF data for applicants, using an analytic technique which he called
"iterative distillation", rather than factor analysis, as others had done,
to examine structure. On the basis of his analyses, Dailey redefined the
second-order anxiety factor slightly and produced a short and efficient
way to determine whether an individual would fall above the cut-off score
on the second-order anxiety factor if the 16PF were scored normally. This
section discusses Dailey's contribution in some detail.

Profiles. In 1971, Dailey (Note 3) compiled average 16PF profiles
over the 16 primary factors, for 25 ATC and related groups. The ATC groups
were: center personnel (n = 6231) and tower personnel (n = 4816), both
tested in 1966; applicants (n = 396); training failures (n = 173); incum-
bent controllers who had been working for one year (n = 140); and a sub-
group of incumbents who had extensive air traffic control experience be-
fore being selected by the FAA and, therefore, were assumed to be extreme-
ly highly qualified (n = 27). Related groups included: two groups of sky
marshals, one tested at Fort Dix (n = 102) and the other tested under civil
service conditions (n = 100); a group of test pilots (n = 30); and a group
FSS incumbents tested on the job (n = 16). Additional groups were obtained
from the manual for the 16PF; these included the standardization sample and
what Dailey called the Faker Group. This last group consisted of 163 col-
lege students who were instructed to take the test under conditions of
faking the best possible and most favorable impression that they could make
on the test.

The profiles for the ATC center personnel and the ATC tower personnel
were very similar. Also, the profiles of all the ATC groups that had taken
the 16PF as applicants were very similar to each other. There were slight

differences between the profiles of the two ATC groups tested while already employed and the four ATC groups tested as applicants. The applicant groups were higher in emotional stability (higher C), more venturesome (higher H), more practical and careful (lower M), less apprehensive (lower O), and lower in free-floating anxiety or less tense (lower Q_4) than were the employed groups.

Dailey, noting that the ATC applicant groups were more like the Faker Group than were the ATC employed groups, correlated the profiles of the 25 groups. He found extremely high correlations among test pilots, FSS personnel, ATC failures, ATC applicants, ATC incumbents, and sky marshals tested under civil service conditions. The average correlation among the profiles for these groups was .971. All of these profiles correlated very highly with the Faker Group (Average correlation: .912). On the other hand, the profiles of the ATC tower and center groups had much lower correlations with the profile of the Faker Group (.72 and .67, respectively). In addition, the profile of the Fort Dix sky marshals correlated .77 with the profile of the Faker Group, and resembled those of the ATC tower and center personnel, tested under experimental conditions, more than the profiles of the sky marshals tested under civil service, pre-employment conditions (correlations of .93, .95, and .84, respectively). Dailey concluded that the air traffic controllers tested under competitive conditions as a prerequisite for employment may have distorted their responses appreciably in the favorable direction.

Dailey also performed a factor analysis on the correlation matrix for the 25 groups and extracted three factors. The first factor he labeled the Paragon Factor or the Social Desirability Factor. Groups with high rotated factor loadings on this factor were ATC applicants, ATC failures, sky marshals tested under civil service conditions, test pilots, ATC incumbents, ATC incumbents with prior controller experience, and the Faker Group. Dailey labeled the second factor the General Public Factor. This included the standardization group, business executives, industrial plant foremen, accountants, store managers, and policemen. The third factor was ill-defined and included artists and physicists. The loadings for the ATC center group and the ATC tower group were evenly split between the first and second factors. Dailey concluded that this analysis demonstrated the influence of the mind set of the individual taking the 16PF on the results obtained. As noted previously in this chapter, this has generally been regarded as a serious limitation of self-report personality inventories. Dailey called for the development of a fake-resistant scoring system that would be usable under competitive conditions, in which individuals take the 16PF as a condition of being hired.

Iterative Distillation. Dissatisfied with previous attempts to demonstrate the factorial validity of the 16PF, and convinced that factor analysis is not a suitable approach for grouping items such as those found in the 16PF, Dailey (Note 4) analyzed item responses from the 16PF using a methodology for item clustering based on the work of DuBois, Loevinger, and Gleser (1952), called homogeneous keying (see also Loevinger, Gleser

335.

& DuBois, 1953). Dailey termed the process "iterative distillation". He employed this technique to obtain an acceptable set of item clusters which had maximum independence and homogeneity.

An important difference between factor analysis and homogeneous keying is that the former seeks to minimize the number of dimensions needed for a structure, while the latter seeks to maximize the number of independent scales obtained from a pool of items. DuBois et al. (1952) based their method on the analysis of the inter-item covariances. Each scale (key) is constructed by adding items one at a time to a nucleus of three items, so as to maximize the saturation with respect to the matrix of items from which the item is drawn. The saturation of the test was defined as the proportion of the total test variance attributable to inter-item co-variances, and is a function of coefficent alpha (KR-20). A cycling process involving elimination and addition of items is followed to assure homogeneity and independence of scales.

Dailey first correlated each item in Form A with each of the 16 factor scores in Form B, using data for ATC applicants. Many correlations between items and factor scores in the range of .30 to .59 resulted from this first step. On the basis of this analysis, seven item clusters were identified and seven additional Form A variables were formed. Next, each item in Form B was correlated with each of the 16 factor scores in Form A as well as the seven new cluster scores. The items in Form B were then clustered in accordance with their correlations with Form A; this resulted in a set of homogeneous clusters. In the next iteration, each item in Form A was correlated with the revised cluster scores in Form B and a second cycle set of homogeneous clusters was then developed for the items in Form A. Once again Form B items were correlated with the revised Form A clusters. The iterations continued until the clusters for each Form were relatively homogeneous and independent. Dailey then performed a series of factor analyses of the items in each cluster plus additional items that seemed to have a chance of belonging to the cluster, and extracted a single factor for each cluster. Items were dropped or added as appropriate, until clusters emerged with the desired characteristics.

At the same time, Dailey (Note 5) continued to examine the potential of the profile scoring of the 16PF; apparently, he became increasingly convinced of its shortcomings when administered under competitive conditions. He cited the work of Sells et al. (1968), as evidence of the inadequacy of Cattell's use of factor analysis as a means of establishing the structure of the 16PF. However, while he argued against the clinical interpretation of the full 16PF profile, Dailey did indicate that the 16PF appeared to measure anxiety or general adjustment rather well. The second-order anxiety factor did seem to display the properties of homogeneity and independence that he was attempting to achieve through the use of iterative distillation. Karson (Note 6) commenting on one of Dailey's papers (Note 7), argued for the value of a profile-producing instrument such as the 16PF, when used by a clinician. In rebuttal to Karson, Dailey (Note 8) noted the extremely low anxiety scores for a group of recent

336.

applicants, where 92% scored at or below 5 and 51% scored 1 or 2. Dailey also noted that the motivation distortion scores for this group were high, with 35% at 10 or greater. Dailey's point was that the 16PF, when administered under competitive conditions, was highly fakable, and for that reason not useful for profile interpretation. Nevertheless, despite the highly transparent nature of the instrument, some applicants did score above the cut-off score of 9.5 on the anxiety scale, and Dailey believed that these applicants should be identified.

The results of his analyses, and the work of other researchers, in particular, Eysenck (1972), Howarth and Browne (1971), and Sells et al. (1968), led Dailey to abandon the attempt to produce homogeneous and independent scales for the entire 16PF, and to concentrate his efforts on improving the measurement of the second-order anxiety factor. Dailey directed his work on iterative distillation toward this end, and this resulted in the identification of 38 items, 18 on Form A and 20 on Form B, which had high correlations with the second-order anxiety factor. These 38 items formed the basis for his revised scale and scoring system.

Revised Scoring System. The simplified scoring system developed by Dailey is based on the 38 items that contain the basic rudiments of the anxiety or neuroticism score. The main difference between the simplified score and the full-scale anxiety score is that items from Factor H (Shy vs Venturesome) were excluded. According to Dailey (Note 9), the simplified score has the advantage of measuring essentially the same trait as the second-order anxiety factor, and it is ". . . a purer measure less subject to chance fluctuation." Dailey (Note 10) reported a correlation of .78 between the simplified score and the full-scale second-order anxiety score.

The simplified scoring proceeds as follows. First, the 18 items on Form A are scored. The most "positive", "favorable", or "healthy" response is allotted one point, so the scores can range from 0 to 18. The majority of respondents score 10 or above on Form A. If this occurs, no further scoring is required and the individual is cleared, since 10 or above on these items is equivalent to at most a score of 3 on the full-scale second-order anxiety factor. If the score on these 18 Form A items is less than 10 for an individual, the 20 items on Form B are scored, allotting one point as was done with the Form A items. If the combined score from Form A and Form B is at least 10, no further scoring is done and the individual is cleared, since a combined score of at least 10 is equivalent to a full-scale anxiety score of at most 5. About 90% of the combined scores are 20 or more. In short, a candidate is cleared as soon as a score of 10 is reached when scoring Forms A and B.

If a respondent has a combined score of less than 10, then 13 of the 36 items, six on Form A and seven on Form B, are rescored. Four of the six items on Form A are from Factor Q_4; the other two are from Factor Q_3. On Form B, three items are from Factor Q_4, two items are from Factor O, and one item each is from Factor C and Factor N. On the

337.

rescoring of these 13 items, one point is allotted for the choice of the "neutral" alternative, and two points are allotted for the choice of the most "negative", "unfavorable", or "unhealthy" alternative. Thus, scores on the rescoring can range from 0 to 26. An applicant is not cleared if one of the combination of scores as given in Table 2 is obtained. Approximately one percent of the applicants will not be cleared using this simplified scoring system.

Phase III

The use of the simplified scoring system based on the 38 core items became operational as a case-finder method for all ATC applicants in May of 1975. Part of the rationale for developing a simplified scoring procedure was to meet the time constraints imposed by the hiring process. With the use of templates, a clearing score can be obtained for most applicants in a matter of seconds. The simplified scoring system provides a quick and efficient way of identifying applicants requiring referral for psychiatric evaluation.

To preserve security, all answer sheets are scored at the Office of Aviation Medicine in Washington, D. C. If an applicant is not cleared, the appropriate region is notified by telephone. Later, a complete listing of all applicants, including those not cleared on the 16PF, is sent to the appropriate region.

Applicants who are not cleared are interviewed by a psychologist or psychiatrist, and usually given a complete examination. A telephone survey (12/14/81) of six of the eleven regional air surgeon's offices indicated that the prescribed procedure is followed virtually in every case, and that, if an applicant does not pass the complete psychiatric examination, that individual is either not hired or not retained if already in training. The respondents to the telephone survey, usually the Regional Flight Surgeon, indicated that most applicants not cleared on the 16PF pass the psychiatric examination. The failure rate seems to vary between 10% and 50%.

Apart from Dailey's reported correlation of .78 (Note 10), there is no formal evidence for the validity of the simplified scoring procedure. Table 3 gives the distribution of applicants who took the 16PF and those not cleared, from 1974 to 1979, according to records in the Office of Aviation Medicine. Since the new scoring system was implemented in May, 1975, the years 1976 to 1979 provide a representative estimate of the percentage of applicants not cleared by the new scoring system. During these combined years, 0.97% of the applicants were not cleared. This number is extremely close to the expected value of one percent if the full-scale second-order anxiety factor had been used. More formal evidence is needed in order to establish the validity of the simplified scoring system.

338.

Table 2

Scoring Combinations on the 16PF for Which

ATC Applicants are not Cleared

Combined A & B (38 items)	Rescored A & B (13 items)
8, 9	17 or over
7	16 or over
4, 5, 6	15 or over
0, 1, 2, 3	14 or over

339.

Table 3

Distribution of ATC Applicants Not

Cleared on the 16PF: 1974-1979

Year	Applicants	Not Cleared	Percent
1974	3,545	128	3.61
1975	2,695	28	1.04
1976	2,975	18	0.61
1977	3,178	33	1.04
1978	3,112	27	0.87
1979	3,372	45	1.33

THE PERSONALITY PROFILE OF THE ATC

Notwithstanding Dailey's analyses, the 16PF data for ATCSs have been subjected to many analyses and profile interpretations. The first profile interpretation.was by Siegel (1966), who reported statistically significant differences on the mean scores of 13 of the 16 primary factors, when comparing the profiles of the ATC center personnel with those of the standardization sample. The controllers were higher on intelligence (B); more exacting, precise, and controlled (Q_3); tougher and more realistic (I); more reliable and conscientious (G); less anxiously insecure (L); more practical and conventional (M); less critical and rebellious (Q_1); lower in free-floating anxiety (Q_4); higher on ego strength or frustration tolerance (C); less guilt prone and worried (O); more reserved and aloof (A); more independently self-sufficient in their work habits (Q_2); less bold and adventurous and more reactive to external threat (E) than the male standardization sample. Siegel concluded that, as a group, the center controllers possessed higher intelligence, greater self-discipline and self-control, tougher realism, greater conscientiousness, and less anxious insecurity than the general population.

Karson and O'Dell (1974b) compared the 16PF profiles for male and female ATC applicants. Only about two percent of the applicants at that time were female, so only 217 females were included in their study. While 10 of the 17 comparisons (motivation distortion was included) were significantly different, most were not of practical significance; the larger number of males in the study (n = 9,886) caused very small differences to be statistically significant. However, females' scores differed by at least one standard deviation from those of the males, both higher, on sensitivity (I) and imaginative vs. practical (M). Otherwise, males and females had basically the same personality profile. In a related study, Karson and O'Dell (1974c) compared the profiles of the 9,886 male applicants with the profiles of the ATC tower and ATC center personnel, gathered in 1966. Their results were similar to Siegel's (1966) and Dailey's (Note 3). All three studies used the same ATC center data, and Dailey and Karson used the same ATC tower data, but different applicant data. The similarity of these results to Dailey's indicated that the personality profiles of applicants had not changed very much in the intervening years. Generally, the profiles of the applicant group indicated that they were more emotionally stable (higher C), more venturesome (higher H), more practical and careful (lower M), less apprehensive (lower O), and lower in free-floating anxiety or less tense (lower Q_4) than were the tower or center controllers. Karson and O'Dell (1974c) concluded that applicants showed even greater mental health than those already employed. However, they also noted higher motivation distortion among applicants. As Dailey had previously (Note 3), Karson and O'Dell noted the high possibility of faking: "After all, it should not be surprising that people who apply for ATC positions might be expected to present a more favorable picture of themselves than center and tower controllers who are already employed" (p. 1004).

341.

Rose et al. (1978) used the 16PF and the CPI, as well as biographical data, in a major longitudinal investigation of ATCS health changes. Approximately 400 ATCSs from New York and New England were studied over a three year period. During the course of the study, slightly over half of these controllers experienced one or more significant psychiatric problems; however, only a few had chronic problems throughout the study. On the 16PF, Rose et al. obtained profiles similar to those of previous researchers (Siegel, 1966; Karson & O'Dell, 1971, 1974c). On the CPI, the controllers generally scored within the normal range, based on a standardization sample of 5,000 males.

In summary, the general finding of these studies was that the personality profiles on the 16PF of various ATC groups were similar; taken at face value, the profiles of the applicant groups showed slightly better mental health, but this is questioned by the motivational distortion scores. The personality profiles of the ATC groups did not differ appreciably from the profile of the standardization sample. Where specific differences did emerge, they were in the direction of better mental health for the ATC groups.

THE VALIDITY OF THE 16PF FOR THE ATCS

The validity of a test is dependent on the uses to which it is put; tests do not have validity in the abstract. The information sought from a test, the population to which it is administered, and the criterion measures used, all affect the interpretation of its validity. Of the major types of validity-content, construct, and criterion-related or predictive - the latter appears to be most germane to the use of the 16PF with ATCS. In this context, there are at least two relevant criterion-related validity questions. First, to what extent does the simplified score correctly identify individuals who should be referred for psychiatric evaluation? Second, does the 16PF profile information predict the on-the-job success of an ATC? To date, there has been no formal investigation of the first question and very little examination of the second.

In an attempt to gather evidence for the validity of the 16PF in the occupational setting of the ATCS, Karson (1969) related peer and supervisory ratings of job performance to the personality factors measured by the 16PF. The ratings were highly related to each other, but no relationship was found between the ratings and the personality variables. Karson and O'Dell (1971), using a different estimate of job performance, the FAA's Employee Appraisal Record for Nonsupervisory Employees, found only one of the 16 correlations (factor Q_2) to be significant. Karson and O'Dell indicated that such low correlations might be due to the restricted variance of the supervisory ratings. Nevertheless, they were not able to present acceptable evidence for the predictive validity of the 16PF with respect to on-the-job performance.

Colmen (1977) examined the validity of the 16PF for the selection and placement of ATCSs, using data from over 2,000 controllers, employed and separated, from 1969 to 1976. The criterion measures used by Colmen

342.

were performance on laboratory exercises in the FAA academy, supervisor ratings, progress since employment, attrition, and an aggregate of the four. Colmen found low correlations between the 16PF and the criterion measures, but concluded that the 16PF did not add sufficiently to the cognitive predictors already found to be valid to justify its additional cost in selecting ATCSs. However, Colmen did find higher correlations when the data were analyzed by specific ATC options (FSS, VFR, IFR, ARTCC). He concluded that the 16PF may have some utility for placement decisions into specific ATC options. It is possible, however, that the correlations Colmen obtained changed simply as a function of the way his data were aggregated. Colmen did not study the validity of the 16PF for use strictly as a case finder, to identify individuals to be referred for psychiatric evaluation.

While no formal research has been conducted on the validity of the case finder utilization of the 16PF, Rose et al. (1978) presented some indirect evidence concerning this. Rose found that those who scored low on the CPI on sense of well-being, responsibility, socialization, self-control, tolerance, and intellectual efficiency were at much higher risk for psychiatric health change, and had a much higher rate of medical disqualification from the FAA for either psychiatric or medical conditions. Trites, Kurek, and Cobb (1967), in an earlier study of 338 ATCSs, found that the difference between underachievers and overachievers at the FAA Academy, as predicted by aptitude scores and biographical characteristics, was a function of essentially the same CPI scales. Rose suggested that the CPI be added to the screening battery for the pre-employment selection of ATCSs, or, if that were not desirable or feasible, that scores from the CPI be related to scores from the 16PF, using his data. This suggestion has not been implemented. Campbell and Chun (1977) found considerable overlap between the CPI and the 16PF at both the scale and second-order factor level, and especially for anxiety (adjustment) and introversion-extraversion. At least by implication, these data provide some suggestive evidence for the validity of the 16PF as a case finder for ATC applicants. More formal validation studies are needed for this case finder use of the 16PF.

LIFE HISTORY DATA

The purpose of a preemployment screening program is to predict the on-the-job behavior of applicants after they have been employed. One source of predictors that has been used for years by industrial psychologists is personal history data provided by applicants on questionnaires or application blanks (Guion, 1965). Another potential source of predictors for ATC applicants is information obtained from background and reference checks that are required of all applicants. Underlying the use of life history data as predictors having some utility for selection is the assumption that experience "programs" the individual and, hence, future behavior is to a great extent a projection of his or her past behavior in similar situations. This assumption is relevant for ATC applicants, since many applicants have had prior ATC experience in the military; however, the focus here is on personality variables.

343.

Usually the information sought in a biographical data form or application blank is factual, but items tapping values, attitudes, and personality characteristics are often included. Sundberg (1977) identified three types of information that are usually obtained: demographic (age, sex, birth place, marital status, etc.); experiential (educational history, work history, military record, etc.); and behavioral (hobbies, leisure activities, distance travel to work, etc.). Such information is often included in traditional application blanks and also in questionnaire forms which utilize standardized multiple choice questions. Information from a traditional application blank is generally used in an informal, non-empirical way, while information from a typical biographical data questionnaire can be used according to results of criterion-related validity studies. On the basis of such studies, items are given weights in relation to their correlation with the criterion.

Background checks involve the verification of information concerning the applicant, as well as the determination of additional information not provided by the applicant on the application blank. Reference checks ordinarily involve the obtaining of character statements from persons whose names have been provided by the applicant. The major sources of information for background and reference checks are letters of recommendation, contact with previous associates and employers, review of military records, and field investigations.

More research has been conducted on the validity of biographical data than on the validity of background and reference checks. Many studies have reported positive findings with weighted application blanks (Frank et al., 1981). Guilford and Lacey (1947) reported validity coefficients between .35 and .40 for weighted application blanks in the prediction of success in training Air Force student pilots, and of .25 to .30 in the prediction of success in training Air Force student pilots, and of .25 to .30 in the prediction of navigator training success. Owens (1976) has reported that correlations between selected information obtained from weighted applications blanks and success in employment usually range from .40 to .60. Potkay (1973), in a review of case history information, concluded that ". . . personal history data as a source of clinically descriptive or predictive information is at least as effective as information derived from psychological test sources" (p. 208).

There are some problems associated with the acquisition and use of life history data. Sundberg (1977) pointed out that life history assessment suffers from inadequate development of conceptual framework and methodology. Information can be falsified, forgotten, or distorted unintentionally. Furthermore, changes in an individual's circumstances or in the social climate can affect the predictive validity of this information (Guion, 1965). Despite these weaknesses, Sundberg (1977) concluded that biographical data have generally been powerful predictors of future behavior.

ATCS research using life history data has been limited. Colmen (1977) included a biographical inventory along with the 16PF to examine their effectiveness as predictors of ATCS success measures. The biographical

344.

inventory consisted of 58 background data items, such as prior experience, performance in school, work history in jobs involving stress, shift work, verbal fluency, and the like. Five scores were derived from combinations of items keyed empirically on 1971 data against attrition, effect on safety, a set of items based on an FSS construct, supervisory or non-supervisory status, and a set of interpersonal items. Colmen found that each of the five biographical inventory scales predicted training performance of both terminal and ARTCC trainees. However, when these predictors were used in conjunction with the 16PF and a semantic differential concept adjective test, little increase in prediction results. It should be noted that Colmen always entered the 16PF information into the regression equations prior to the biographical information. Generally, the first variables entered into a regression equation contribute more to an increase in prediction than those variables which are entered later. For placement decisions, however, Colmen found that the biographical inventory scales did provide additional information for the classification of newly hired ATCSs into individual ATC options, with training performance as the criterion.

Rose et al. (1978) used biographical data in their study of ATCS health changes. They found that applicants over 30 years of age and without prior ATC experience had a higher probability than other applicants of experiencing some psychiatric problem. Also, high use of alcohol, especially at earlier ages, was associated with future difficulty. Finally, a history of legal involvements (being arrested, sued, etc.) was associated with impulse control disorders, psychiatric problems, and eventual disqualification. The authors suggested that data concerning the above predictors might be included in the biographical history information requested prior to employment.

VanDeventer (1980) examined biographical data on ATC trainees who entered the Academy between June 1977 and June 1979, to determine the relationship of these data to training success or failure. The biographical questionnaire consisted of 60 items covering high school education, education beyond high school, and prior experience. Complete data were obtained on 2,371 trainees, 1,257 with prior ATC experience, usually in the military. The results showed that pre-FAA ATC experience was a significant factor in training success, and that education was not related to success. These findings confirmed previous ATCS studies (Cobb & Nelson 1974: Cobb, Young, & Rizzuiti, 1976). However, they did not identify personality variables.

At present, a project in the Southern Region of the FAA is seeking to improve the ATCS selection procedures, using information from a life experience questionnaire. In addition to examining such variables as education and prior experience, this questionnaire includes information on interests, motivation, and other personality traits. The rationale for including the latter kind of variables is that they can have an impact on job behavior and, consequently, on job success. To date, no reports from the project are available.

Personality assessment of ATC applicants is included by most countries as part of their selection procedures, principally based on written tests and selection interviews. Normally, the latter are not clinical interviews as such.

The written personality assessments presently used by countries other than the United States for ATC applicants apparently have been developed in-house, usually as part of a larger selection battery. Major development of selection tests has been done at the Rijks Psychologische Dienst (R.P.D.), the Institute of Psychology of the Netherlands Government at the Hague. In 969, the Director General of Eurocontrol requested the R.P.D. to assist in the selection of ATC trainees, and since then formal assessment for Euro-Control has been done by R.P.D. (Note 11). Prior to 1969, R.P.D. had had several years experience in the selection of Dutch ATCSs. The selection tests developed by R.P.D. include some personality subtests. Besides assessing the candidate's mental functions, such as logical deduction, analytical thinking, ability to abstract, numerical ability, and rapidity and accuracy of work, the psychological battery assesses personality characteristics, sich as ambition, perseverance, adaptability, and stability.

Until recently (Note 12), an adaptation of the 16PF was used in Australia as part of an ATCS selection battery. Only 45 items, which had previously been identified as providing a valid assessment of applicants, were scored. The Australians continue to use a map reading test, interrupted at irregular intervals, as a measure of short memory as well as tolerance to a mild degree of stress. The United Kingdom has experimented with the 16PF, with volunteers from its CAA staff (Note 13), but apparently the 16PF has not been used with applicants. It is interesting that the first analysis of the 16PF results indicated that controllers in the United Kingdom had a markedly different profile from controllers in the United States. No specifics were given concerning these differences.

Written personality tests are also used for ATC applicants in Sweden and Japan. Applicants to the Swedish Air Traffic Services Academy are tested prior to admission. Oloffson (1980) described traits that an applicant must possess, but he did not indicate the tests used to assess them. The traits mentioned by Oloffson include the ability to withstand stress, the ability to accept responsibility, and psychological balance. Maekawa (1980) reported that an entrance examination is required for admission to the Aeronautical Safety College, the controller education facility of the government of Japan. Apparently, screening tests developed under the auspices of medical personnel are used as part of the selection process.

Most countries use a selection interview for ATC applicants, one purpose of which is to assess the personality of the applicant. In Norway (Note 14), members of the ATS administration staff serve on a selection

board. In particular, knowledge of English, job motivation, and mental fitness of candidates are assessed. In the Netherlands (Note 15), candidates are interviewed first by a psychologist and then by a selection committee. In particular, motivation and suitability are assessed. An interview is also used by the Royal Netherlands Air Force in the selection of controllers (Note 16). The aptitude test results of the candidate are not known by the interviewers, who attempt to assess social adjustment, discipline, sense of reality, independence, and motivation of the candidate. Canadian ATC applicants are interviewed in order to determine their decision-making ability, tolerance for stress, initiative, tenacity, vigilance, and cooperativeness (Note 17). Finally, ATC candidates from the United Kingdom are interviewed by three senior ATCSs (Note 13), and those from Australia by a three-person panel, one of whom is a psychologist (Note 12).

SUMMARY

The 16PF has been used by the FAA for personality assessment of ATC applicants since 1966. The inventory was incorporated into the medical screening of applicants as a case finder to identify individuals who should be referred for psychiatric evaluation. During the first nine years of its use, a full profile scoring of the 16PF was performed for each applicant; decisions concerning further evaluation were based primarily on the second-order anxiety factor. Since May of 1975, decisions concerning further evaluation have been based on the scoring of 38 items identified by John Dailey as measuring essentially the same trait as the second-order anxiety factor. The revised scoring is easily performed, using templates. Approximately one percent of the applicants are screened out (not cleared) by the revised scoring procedure and are referred for a psychiatric evaluation. This percentage is in agreement with the expectations set when the cut-off score was first adopted.

In addition to ATC applicants, the 16PF has occasionally been given to employed controllers. A large-scale testing of ATC tower and ATC center personnel was conducted in 1966 prior to the testing of ATC applicants. While not differing greatly from the 16PF profile of the general public, the 16PF profiles of various ATC groups have shown that controllers have greater self-discipline, higher intelligence, greater conscientiousness, and lower anxiety than the general public. The 16PF profiles of employed controllers and applicants have differed only slightly; applicant profiles appeared to show better mental health, but this is believed to be explained by conditions under which each group was tested.

No formal empirical evidence is available for the validity of the revised scoring of the 16PF, used as a case finder. In an informal survey of six regional medical offices, the Regional Flight Surgeons were of the opinion that most applicants who are not cleared on the 16PF do pass the subsequent psychiatric examination. The estimates of failing to pass the complete examination after not being cleared on the 16PF ranged from 10 percent to 50 percent. The validity of the revised scoring system needs further study.

The relationship between personality inventory scores and job perform-
ance measures typically has been quite low, for most occupations. There
is little evidence for the utility of a measure such as the 16PF in the
prediction of on-the-job success. In fact, the use of personality inven-
tories as the basis for job selection decisions was prohibited by the U.S.
Civil Service Commission in 1965 mainly because of the poor validity of
these inventories in predicting job success. This prohibition continues
under the present Uniform Guidelines. However, the use of such inventories
to identify individuals who may have personality disorders so severe as to
impede their ability to perform according to minimum job standards is con-
sidered appropriate and needed. Hopkin (1980), in a review of the litera-
ture on measurement of air traffic controllers, suggested that it may be
necessary to understand controllers first as humans and only second as con-
trollers. While Hopkin indicated that the ". . . relevance of personality
remains unclear" (p. 554), he did see relevance in attempting to measure
proneness to anxiety, as with the 16PF, as a possible way of predicting the
difficulty that a controller may have in facing the responsibility of con-
trolling traffic. The use of a personality inventory as a case finder is
not prohibited by the Uniform Guidelines. If a suspected severe person-
ality disorder is confirmed by psychiatric evaluation, then that individual
can, and should, be denied a job as an air traffic controller.

The identification of potential personality disorders which could in-
hibit the performance of an air traffic controller could also be accom-
plished by interview. Selection interviews are routinely used by many
countries in screening ATC applicants. In the United States, selection
interviews are not routinely used primarily because of the large number
of applicants, but also, undoubtedly, because of the lack of demonstrated
validity for such interviews. Clinical interviews are used by the FAA
only for those ATC applicants who are not cleared on the 16PF.

Another source of information that could be helpful in screening ATC
applicants for personality disorders is life history data obtained through
the use of application blanks and background and reference checks. Life
history data are gathered routinely for ATC applicants, but to date these
data have not been alalyzed systematically. Research with life history
data in other occupations has been promising. There has been little formal
work conducted on the utility of this information for ATC applicants. It
is evident that prior ATC experience is very helpful, but not very much is
known about the utility of the personality and behavioral information that
can be obtained.

In summary, further research in the area of personality assessment
and screening of ATC applicants appears to be indicated. Issues identi-
fied in this paper include:

1. Examination of the validity of the revised scoring of the 16PF,
    used as a case finder;

348.

2. Examination of the potential of life history data as a personality case finder similar to the present use of the 16PF;

3. Examination of the relationship between life history data, personality variables, and on-the-job performance.

349.

REFERENCE NOTES

1. Reighard, H. L. Psychiatric assessment of air traffic controllers. Unpublished manuscript dated June 5, 1969.

2. McKee, W. F. Letter to Mr. John W. Macy, Chairman, U. S. Civil Service Commission. January 5, 1967.

3. Dailey, J. T. Motivation distortion on the 16PF test. Memo, Office of Aviation Medicine, Washington, D. C., September 30, 1971. This memo includes extensive tables involving profile comparisons, as well as the full correlation matrix and the matrix of factor loadings.

4. Dailey, J. T. An item factor analysis of the 16PF. Memo to AM-300. Office of Aviation Medicine, Washington, D. C., November 1, 1971.

5. Dailey, J. T. Analyses of "Taxonomic investigation of personality. Conjoint factor structure of Guilford & Cattell" by Saul Sells, August 1968. Memo to AM-300. Office of Aviation Medicine, Washington, D. C., November 9, 1971.

6. Karson, S. Comments on Dr. Dailey's memo of August 11, 1972. Memo to Dr. Haynes, August 21, 1972.

7. Dailey, J. T. Comments on "The 16PF and basic personality structure: A reply to Eysenck." Memo to AAM-300. Office of Aviation Medicine, Washington, D. C., August 11, 1972.

8. Dailey, J. T. Criticisms of the 16PF. Memo to AAM-300. Office of Aviation Medicine, Washington, D. C., August 29, 1972.

9. Dailey, J. T. Psychometric assessment of the 16PF test. Memo to Chief, Behavioral Sciences Division, AAM-300. Office of of Aviation Medicine, Washington, D. C., December 29, 1975.

10. Dailey, J. T. New Scoring System for the 16PF. Memo to the Federal Air Surgeon, AAM-1. Office of Aviation Medicine, Washington, D. C., January 7, 1977.

11. Meij, G. J. The selection of Eurocontrol trainee air traffic controllers. Unpublished and undated manuscript.

12. Bartlett, B. Department of Transport, Commonwealth of Australia in a letter to John J. Convey, January 29, 1982.

13. Reply from the United Kingdom to S 2/TG/77 of 24 January 1977, dated February 16, 1977.

14. Replies to "Questionnaire on On-The-Job Training (O.J.T.)," undated memo.

350.

REFERENCE NOTES (Continued)

15. Memo from the A.T.S. and Telecommunication Directorate Holland in reply to Questionnaire S2/TG/76, undated.

16. Rameckers, F. H. J. I. Reply of Royal Netherlands Air Force to Questionnaire on Selection of Air Traffic Controllers (Ref. S2/TG/76 dated 3.7.76), undated.

17. Notes from a telephone conversation of E. Pickrel, Office of Aviation Medicine and D. Kirby, Ottawa, Canada, November, 1981.

351.

Chapter 18

VALIDATION OF NEW ATCS SELECTION TESTS
ON TRAINEE AND CONTROLLER POPULATIONS

THREE STUDIES - 1972, 1977, 1979[1]

This chapter presents summaries of the three major validation studies
that culminated in the formation of the new test battery for air traffic
controller selection that was adopted for operational use by the OPM in
October, 1981. The first two were carried out under FAA contracts by a
group at Education and Public Affairs, (EPA), a private research firm,
led by Joseph G. Colmen. The 1972 study, by Milne and Colmen, was a
concurrent validation of a diverse battery of paper and pencil aptitude
and personality tests and psychomotor tests administered to 800 newly
appointed and experienced controllers, using confidential supervisor
ratings as criteria. The 1977 study, by Mies, Colmen, and Domenech, had
similar objectives but was expanded in respect to sample size, structure,
and issues addressed.

The third study, by James O. Boone (1979a), was carried out at the
FAA Civil Aeromedical Institute (CAMI) in Oklahoma City and utilized a
final sample of 1827 ATC trainees who attended the FAA Academy during
1976 to 1978. This study built on the progress realized in the FAA stu-
dies and identified the final test battery that was further validated and
evaluated in the research presented in Chapter 21.

I

1972 - SELECTION OF AIR TRAFFIC CONTROL SPECIALISTS

Anne Milne and Joseph G. Colmen

Objectives

The focus of this research was on selection as related to job perform-
ance. It was based on concurrent validity analysis. The primary questions
addressed were:

(1) To what extent is it possible to predict job performance
of a journeyman ATCS based on a battery of tests admini-
stered at the time of job application?

(2) To what extent can improvement in prediction be achieved
by selective assignment of applicants to selected ATC op-
tions (FSS, Center, IFR or VFR terminals) and within these,
to high or low density (activity) facilities?

[1]Prepared by S. B. Sells

(3) To what extent do the measures selected to accomplish the
    above objectives affect black and white applicants with an
    equal degree of fairness?

Sample Description

The total sample comprised approximately 800 employees who were either
journeymen ATC specialists (FPL) or new ATC appointees and included an
oversample of black ATC specialists. Sample selection provided for distri-
bution between four types of ATC facilities (Centers, IFR and VFR Termi-
nals, and Flight Service Stations). Samples for Centers and IFR termi-
nals were further stratified between high and low activity facilities.
Race-ethnic distribution, including the oversample, was 93% white and
7% black.

The sample was drawn from FAA facilities in 15 major cities in the
United States and was randomly selected (except for the oversample) from
ATC facilities within a 100-mile radius of the 15 hub cities. Sampling
was controlled to excluse ATC specialists over 36 years of age and to
insure that FPL Specialists had no less than 3 years nor more than 10
years of ATC experience with FAA. Because of the small number of women
in the ATC workforce, it was not possible to stratify the sample selection
based on sex. Participation on the part of ATC specialists was voluntary.
All subjects were administered a battery of paper-pencil tests or forms
including the CSC test battery.

In addition, a subsample of about 260 ATC Specialists, including
journeymen and new appointees, who took the paper-pencil battery, were
administered a series of "psychomotor" tests in a separate testing ses-
sion at FAA facilities at Oklahoma City. Selection of this subsample was
controlled for region, type of ATC facility, and variance on the confi-
dential supervisory performance evaluation which was used as the criterion
meausre. Race-ethnic distribution of the "psychomotor" subsample was 83%
white and 17% black.

Predictors

In selecting predictors for this study, four objectives were consid-
ered. These were: (1) They should cover as broadly as possible the
range of job and worker attributes identified with the ATC occupation;
(2) Experimental tests should not substantially overlap areas already
covered by the existing CSC test battery, which was to be administered
to the ATC Specialists; (3) Experimental tests should be selected which,
based on prior research, appeared to have potential validity for ATC
selection; and (4) The two selected test batteries (paper-pencil and psy-
chomotor) should not require more than 8 hours each to administer. Within
these objectives the following predictor tests were selected:

Paper-Pencil Battery

Aptitude Tests - CSC test battery (5 parts)
 Minimum Coins Test
 Dial and Table Reading Test

Knowledge and Interest Tests - Dailey Technical and
 Scholastic Test
 ATC General Information Test

Personality Tests - Concept-Adjective Test
 Closure Speed Test

Background Information - Biographical Inventory

Psychomotor Battery - Directional Headings Test
 Hidden Patterns Test
 Press Test
 Controller Decisions Evaluation (CODE) Test
 Multiple Task Performance Test
 Compressed Speech Test

The Paper-Pencil Battery was administrated to the total sample group. Test administrators were FAA employees, trained by staff members from Education and Public Affairs. The Psychomotor Battery was given to the subsample at FAA facilities in Oklahoma City, where equipment required for administration was available. Detailed discussion on the source, nature of the test, method of scoring, reliability and validity for the selected tests is provided in Milne and Colmen (1972). The contractor obtained and kept all test results and information on individual ATC Specialists on a confidential basis. No individual data were provided to FAA.

Criterion Measure

After examining a number of alternatives, a Confidential Supervisory Evaluation form was used as the criterion against which experimental tests were validated. The evaluations were obtained by the contractor directly from each ATC Specialists' supervisor. They were not reviewed by FAA nor were copies provided to the agency. The contractor retained all individual evaluation data on a confidential basis. The evaluation covered the broad areas of performance within a number of task behaviors for each area.

Performance Area	Number of Items
Knowledge	3
Perception	6
Comprehension	6
Memory	2
Communication	8
Judgement	4
Traffic Management Techniques	4
Performance Work Stress	4
Interpersonnel Skills	3
Other Personnel Skills	5

355.

In addition, an "Overall Performance" category, consisting of four task
items, and a summary evaluation were obtained. The summary evaluation
was a seven-point rating scale with "1" the highest performance and "7"
the lowest.

Descriptive Information

Based on biographic responses from 304 FPL ATC Specialists, the fol-
lowing background and educational information was obtained:

> 98% were men; 2% women
> 96% had prior military service
> 72% claimed prior (military) experience as a controller
> 2% claimed experience as a pilot
> 48% took the CSC test battery for appointment; 10%, more than
>   once before passing
> 97% completed high school
> 25% attended college; less than 1% completed college

Analytical Methodology

The data collected were analyzed by a variety of statistical treat-
ments, including: (1) multiple regression analysis, in which test meas-
ures were correlated with performance evaluation measures to determine
how well they predicted job performance; (2) analysis of variance, for
assessment of differences between groups, such as minority vs non-minority;
journeymen vs new appointees; and, those hired with previous aviation-
related experience compared to those hired without such experience; and
(3) multiple discriminant analysis, in which tests were evaluated in terms
of their ability to maximize placement within option and activity levels.

Results

In summary, this study concluded that:

1. Analysis of the capacity of the tests to predict job performance
of journeymen ATC specialists produced mixed results. The CSC test bat-
tery was marginal in predicting job performance, as measured by super-
visory evaluations. This result, however, was due primarily to the re-
striction of range of the CSC test predictors, which was attributed not
only to the selection process, but also to the attrition that took place
during the developmental training period as the controllers included in
the sample progressed to the journeyman (FPL) level. This study did not
attempt to correct for the range restriction and consequently the predic-
tive relationship of the CSC Tests to the job performance criteria was
not determined. The psychomotor tests developed consistently significant
correlations with supervisory job performance evaluations.

2. Using a combination of paper-pencil and psychomotor tests, it
was possible to improve assignments to the different ATC options, based
on tests results and, with additional tests, to high or low density centers
or IFR terminals.

3. If relevance to the job is the primary factor in determining test acceptability, the tests proposed in this research met that criterion. They also predicted job performance equally well for blacks and whites. However, blacks as a group consistently scored lower on the tests than did whites.

This study provided FAA with valuable information and insights concerning the problems associated with selection and placement of applicants for ATC work. However, action on the research results and recommendations was deferred due to a number of considerations, including the complexity of the specialized equipment required for the psychomotor tests, logistical difficulties in test administration, and complexity in test scoring and ranking of applicants.

II

1977 - SELECTION OF AIR TRAFFIC CONTROL SPECIALISTS

Joanne Marshall Mies, Joseph G. Colmen, and Oakie Domenech

Continuing concern with the rate of ATC trainee attrition and consideration of the reestablishment of centralized ATC training at the FAA Academy resulted in a review of FAA selection and screening policies in December 1974 (FAA, Note 1) and completion of a cost analysis study of alternative strategies (FAA, Note 2), in March 1975. Among the actions resulting from this review, FAA contracted with Education and Public Affairs, Inc., (EPA) for a follow-on analysis of ATC selection tests, in June 1975. This is summarized in the final report of this study (Mies, Colmen, and Domenech, 1977).

Objectives

The objectives of this research were directed to essentially the same concerns as in the previous (1972) study completed by EPA: selection, placement, and fairness. However, the study design was expanded significantly with respect to sample size, structure, and representation; it encompassed more criterion measures of ATC job success, and also incorporated an evaluation of prior aviation-related experience and educational level as predictors of ATC job success.

Longitudinal Analysis of 1972 Experimental Tests

The first effort in this study was to analyze on a longitudinal basis the relationships between the experimental tests administered during 1971 and criterion measures, for the ATC specialists who participated, as of 1975. Validity coefficients of each test score with each criterion measure were calculated. A factor analysis of the tests was also undertaken to determine overlapping of underlying variables. The tests selected from these analyses were then entered into multiple regression analysis to determine the minimum number of tests or test scores that would predict the

maximum proportion of variance in each of the criterion measures. Table 1 identifies those 1971 experimental tests and predictors (X) which, based on their simple correlation with the various criterion measures established for the analysis, warranted further study by means of field validation.

With some modifications, these experimental tests, which had been administered in 1971, became the "core" test battery for the 1977 research. A more detailed discussion of the methodology and results of this longitudinal study is provided in the report by Mies and Colmen (1976).

Sample Description - 1977 Study

A comprehensive sample design was constructed to define the ATC population to ensure a representative sample for three specific "year of hire" groups. These groups represented three ATC career "stages": (1) New Hires (1976); (2) Developmental ATC Specialists with 2 to 3 years of ATC experience in FAA (1973 and 1974); and, (3) ATC Specialists with 2 to 6 years FAA experience at the journeyman (FPL) level (1969 and 1970). In addition to these primary ATCS samples, three additional ATCS samples were included: (4) An oversample of currently employed women and minority ATCSs in the same three "year of hire" groups; (5) ATC Specialists who participated in the 1972 research; and (6) a sample of ATC specialists who were hired during the three time periods sampled, but who had separated from ATC work before reaching FPL status.

Sample selection was constrained to exclude employees who were in ATC staff or supervisory positions or over age 31 at the time they were hired, except in the FSS option.

Sample section for the three primary groups was based on stratified random sampling methods to provide a proportionally representative group of the total constrained ATC universe for each of the four ATC "options" (FSS VFR, IFR, and ARTCC, with respect to both the initial and current option of assignment.

Table 2 identifies the various ATC specialist samples by year group, the total number desired, and the samples actually obtained. Since participation was voluntary, the number invited was expanded where possible to provide for declination and non-responses.

The numbers of women and minorities who volunteered were too small to enable analysis by year group or ATC option. Consequently, the analysis on test fairness combined year and option groups to provide 235 women and 321 minorities. The oversample for women and minorities was used only for analysis of the fairness of the ATC success predictors. The final EPA report (Mies, Colmen, and Domenech, 1977, Chapter VII) provides a complete discussion of the sampling methodology.

Table 1

Predictor and Criterion Measures of ATC "Success" (1975) for Developmental (DEV) and Full Performance Level (FPL) Controllers who Participated in 1971 ATC Research

| | Criterion Measures | | | | | | |
| Predictors | Separation | Progression + Attrition | | Present Option | | 1971 Supv. Assess. | Sup/Staff Position |
	DEV	DEV	FPL	DEV	FPL	FPL[3]	FPL[3]
CODE	X	X	X	X	X	X	-
Dial Reading	X	X	-	-	-	-	-
Dir. Heading	X	-	-	-	-	X	-
Air Traffic Prob.[1]	-	-	-	X	X	X	-
Arith. Reasoning[2]	-	-	-	-	-	-	-
ATC General Info.	X	-	-	-	-	X	-
Concept Adjective	X	X	-	X	-	X	-
Biographical Info.	X	X	-	X	X	X	X

1) Air Traffic Problems, CSC Test No. 540
2) Arithmetic Reasoning was Part 5 of the Dailey Technical and Scholastic Test (TST)
3) Supervisory assessment criterion data available for 1971 FPL ATCS only. No new 1971 appointees progressed to ATC supervisory or staff positions by 1975.

359.

Table 2

ATC Specialist by Year Hired.
Sample Groups and ATC Career Status

Year Hired	ATC Sample Year Groups	ATC Career Status	Number Invited	Desired Sample	Obtained Sample	Percent of Desired Sample
1969-70	Employed ATCS	FPL	1344	800	754	94
	ATCS Oversample	FPL	151	200	31	6
	Separated ATCS	DEV	--	--	362	--
1973-74	Employed ATCS	DEV	1127	800	740	93
	ATCS Oversample	DEV	258	200	72	36
	Separated ATCS	DEV	--	--	166	--
1976	Employed ATCS	New Hires (DEV)	610	610	590	97
1971	Employed ATCS	FPL	480	480	270	56
	Separated ATCS	FPL/DEV	--	--	74	--

360.

Predictors

Two of the tests used in the 1972 study were modified for use in the 1977 research.

CODE (Controller Decision Evaluation). This test consisted of three film versions of a computer simulation of moving air traffic patterns appearing on a radar scope. Initially these were converted to slide projector presentation to eliminate the need for movie projector equipment and to simplify both the response recording and scoring. Group administration of the various adaptations of CODE during this test development phase clearly pointed up the practical problems of using film or slide projection equipment in test administration. Subsequently, a paper-pencil version was developed which incorporated measures of abilities to identify potential conflicts of aircraft as well as the traditional kinds of aptitudes within an air traffic control context. The resulting test, the Multiplex Controller Aptitude Test (MCAT) was made available to EPA as a substitute for the CODE tests. Development of the MCAT is described in Chapter 15.

Arithmetic Reasoning. In the prior research, the arithmetic reasoning test was one part of the Dailey Technical and Scholastic Test (TST). Since this part could not be given as an isolated test, a similar Arithmetic Reasoning test developed by the Army Air Forces, was used in the 1977 research.

In addition to the predictors derived from the longitudinal analysis of the experimental tests administered in 1971, three other predictors were used.

Pre-employment Experience Questionnaire (PEQ). To obtain specific data from participating ATC specialists on various kinds of pre-FAA experience and education, a questionnaire was developed from the CSC Rating Guide elements used as a basis for granting additional credit in the ATC employment selection process.

ATC Occupational Knowledge Test (OKT). This test was developed to be "job-knowledge specific." Consequently, it was not included in the 1977 research for the purpose of evaluating its use in screening ATC job applicants for employment eligibility. It was used to measure the "quality" of prior experience as a potentially improved basis for granting additional credit for experience in the selection process in place of the existing CSC Rating Guide. Development of the OKT is reported in Chapter 16.

Sixteen Personality Factor Questionnaire (16 PF). The 16 PF Questionnaire is administered as part of the medical qualification process to all entering ATC Specialists (See Chapter 17). It is designed to measure personality characteristics not otherwise measured and the 1977 research offered the opportunity to assess its utility for possible selection or placement purposes.

361.

The experimental test battery used in the 1977 research may be divided roughly into two groups of instruments; (1) Cognitive Predictors --consisting of tests that have predetermined right or wrong answers; and (2) Other Predictors -- in which applicant responses ordinarily have no inherent value of right or wrong, except as measurable against criteria or values external to the test. The following identifies the predictors used within these two groups.

Cognitive Predictors

Multiplex Controller Aptitude Test (MCAT)
Directional Headings Test (DHT)
Dial Reading Test (DL-RD)
Arithmetic Reasoning Test
ATC General Information Test
CSC Test Battery (5 parts)
ATC Occupational Knowledge Test (OKT)

Other Predictors

Pre-Employment Experience Questionnaire (PEQ)
Concept Adjective
Biographical Inventory
Sixteen Personality Factor Questionnaire (16 PF)

Rather than re-administer the CSC test battery, the intent was to obtain the CSC test scores for participating ATC specialists from existing records. However, this proved to be infeasible for a large number of the employees in the sample. Consequently, analysis of the tests in the CSC test battery as predictors of ATC success was not possible in this study. The cognitive predictors and other predictors were each analyzed separately against the criterion measures used to define success in the ATC occupation. They were also analyzed together in a supplemental report (Colmen, 1977) to determine the extent to which other predictors added to the validity of the cognitive predictors. Mies, Colmen, and Domenech (1977) provided a discussion and description of each of the predictors used in the study, including relevant prior research, reliability, and validity.

For analytic purposes, the objective was to obtain pre-employment and education information on all ATC specialists in the sample groups. The experimental test battery was administered to approximately 50 percent of the 1969-70 and 1973-1974 samples and to all new hires in the 1976 group. Selection of employees to take the experimental test battery was controlled to provide proportional distribution by ATC option within each year group. The experimental tests were administered to the new hires at the FAA Academy, on their first day of attendance, by the EPA staff assisted by FAA personnel. For those ATC specialists assigned to facilities, the tests were given by FAA test administrators who were trained by EPA.

Criteria and Measures of ATC Success

To determine the validity of experimental tests, prior experience, and education as predictors of ATC success, it was necessary to establish operational definitions (criteria) of success and ways to measure them. The following four criterion measures were used in the study and were then combined into a single aggregate measure of ATC success:

(1) <u>Training Performance</u>. This was measured by scores received on the ATC Laboratory Problems and the Controller Skills Test during initial ATC training at the FAA Academy, which are required of students in order to demonstrate operational application of academic knowledge.

(2) <u>On-the-job Performance</u>. This was measured by confidential job-task assessments prepared by each employee's supervisor. It included 54 questions on ATC job tasks and four general questions on quality of job performance. Responses to a seven-point overall rating scale were selected as the measure for on-the-job performance.

(3) <u>Progression</u>. For each ACTS, this was measured by the ATC option to which he or she was assigned initially when hired, compared to the option to which assigned on January 1, 1976. Four options were defined to represent general complexity levels of ATCS work (FSS, VFR, IFR, ARTCC) with ARTCC defined as the highest complexity level. Within this hierarchy, progression scores were established as follows: High -- assigned when an ATCS was in an option of a complexity level <u>the same as</u> or <u>higher</u> than the initial option to which assigned (i.e., <u>VFR initial</u> assignment; IFR current assignment on January 1, 1976), and <u>Low</u> -- assigned when an ATCS was in an option of a <u>lower</u> complexity level than the option to which initially assigned (i.e., ARTCC initial assignment; FSS current assignment on January 1, 1976).

(4) <u>Attrition</u>. This was measured by whether or not ATC Specialists hired during the year groups sampled for this study were still employed by FAA in ATC work. Those still employed as ATC Specialists were assigned a high score, and those seprated were assigned a low score.

(5) <u>Aggregate Criterion of ATC Success</u>. This was constructed from the four individual criteria (training, on-the-job performance, progression, and attrition) and provided a 5-point scale value of ATC success.

The <u>predictors</u> of ATC success used in this study -- experimental tests, prior experience, and education -- were analyzed for validity against each of the four <u>criterion measures</u> of ATC success -- training scores, supervisory assessments, progression, and attrition. Final conclusions and recommendations were based on the validity of the predictors with the aggregate criterion.

Descriptive Information

Date on the education and experience background of ATC Specialists who participated in the 1977 study are provided in the EPA Final Report.

The most marked changes between the three "year-of-hire" groups (and the 1972 sample) were in educational level, military service, and pilot experience. Table 3 compares these sample groups on selected variables. The sample of 304 ATCS hired prior to 1971 was derived from the 1972 study; these ATC Specialists could have been hired between 1960 and 1970.

Analytical Methodology

Before undertaking the validity analysis, several preliminary analyses were made. These were required (1) to substantiate that the sample of ATC Specialists who did volunteer was not, except for the proportion of minorities, essentially different from those who did not volunteer, (2) to calculate the means, standard deviations, reliability, and inter-correlations of the experimental predictors and ATC success criterion measures, (3) to determine that the order of MCAT test administration did not affect test results, (4) to convert the training scores to standardized (z) scores, (5) to analyze pre-employment experience and establish empirical scale values for validation, and (6) to analyze predictors for inclusion in the validation study.

Detailed information concerning these preliminary analyses, the statistical results on distribution, means, standard deviations, and inter-correlations is provided in the EPA Final Report (Mies, Colmen, and Domenech, 1977). Results of preliminary statistical analyses for the non-cognitive predictors are reported in Colmen (1977). As a result of these preliminary analyses, the following cognitive tests were excluded from the validation analysis: the CSC Battery, which was not available for a significant number of ATC Specialists in the sample and for which the effect of restriction in range on those scores which were available seriously interfered with proceeding with the validity analysis, and the ATC General Information Test, which was administered only to the new hires attending the FAA academy training; it was not given to the 1973-74 or 1969-70 ATC samples because of time constraints on field testing.

Consequently, the following predictors were included in the primary validation analysis:

Multiplex Controller Aptitude Test (MCAT)
Directional Heading Test (DHT)
Dial Reading Test (DL-RD)
Arithmetic Reasoning Test
ATC Occupational Knowledge Test (OKT)
Pre-employment Experience (PEQ)

The following non-cognitive predictors were included in the supplemental validation analysis:

Concept Adjective
Biographical Inventory
Sixteen Personality Factor Questionnaire (16PF)

364.

Table 3

Education and Experience Levels

Year Hired as ATCS

	Prior to 1971 (N=304)	1969 1970 (N=659)	1973 1974 (N=661)	1969-70 1973-74 Oversample (N=103)	1976 (N=592)
High School	72%	34%	24%	25%	16%
Some College	25%	53%	53%	59%	56%
College Degree(s)	1%	13%	23%	16%	28%
Military Service	96%	75%	74%	56%	71%
ATC Experience	72%	--	--	--	--
IFR	--	35%	39%	27%	32%
VFR	--	39%	38%	26%	37%
Pilot	2%	30%	34%	11%	33%

In summary, the validation methodology for the cognitive predictors (and prior experience) consisted of the following steps:

1. Statistical analyses of each of the smallest homogeneous samples (i.e., each ATC option and all ATC options combined within each year group against each of the four individual ATC success criterion separately). This resulted in 118 separate analyses.

2. Predictors selected from Step 1 (based on validity co-efficients and significance levels) for each ATC option and all ATC options combined across all year groups were then analyzed to identify the best overall set of predictors for each of the four individual criterion measures. This resulted in 17 separate analyses.

3. Results for each of the four individual criterion measures were then examined for each ATC option and all options combined to determine the best set of predictors across criterion measures. This resulted in 17 separate analyses.

4. The final sets of predictors from Step 3 were then validated against the aggregate ATC success criterion, leading to derivation of weighted test and experience scores. This involved five separate analyses.

5. The weighted test battery and experience scores were then validated against the four single criterion measures. This involved four separate analyses.

The methodology described above was modified for analyses of the non-cognitive predictors, to compensate for the differences imposed by the use of personality as opposed to aptitude or knowledge tests. The most significant change in methodology resulted from the need to establish empirically the appropriate directionality of the various non-cognitive predictor scales before undertaking validity analysis. The specific analytical techniques applied in evaluating the non-cognitive predictors are identified in the report by Colmen (1977).

The methodology to evaluate the fairness of the experimental test battery, prior experience, and the Occupational Knowledge Test was focused on the determination of differential validity in accordance with the Uniform Guidelines on Employee Selection Procedures (1978) which state:

"When members of one racial, ethnic, or sex group characteristically obtain lower scores on a selection procedure than members of another group, and the differences are not reflected in differences in measures of job performance, use of the selection procedure may unfairly deny opportunities to members of the group that obtains the lower scores."

The approach followed is discussed by Mies, Colmen, and Domenech (1977) in the EPA Final Report.

Results

The following is a brief summary of the results of this research.
More detailed discussion is provided in the EPA reports by Mies and Colmen (1976), Mies, Colmen, and Domenech (1977), and Colmen (1977).

Selection. Two of the experimental tests -- MCAT and Directional
Headings -- predicted the ATC success criteria established for the study
at statistically significant levels of confidence for all ATC options
combined and for the three primary year groups sampled (1969-70; 1973-74;
and 1976). Two other experimental tests -- Arithmetic Reasoning and
Dial Reading -- either did not predict the ATC success criteria or did
not add appreciably to the prediction values obtained from MCAT and Directional Headings.

Preemployment aviation-related experience and the OKT, while not intended for use in determining initial appointment eligibility, predicted
ATC success at statistically significant levels of confidence and increased the validity coefficients obtained with the experimental tests.

Education beyond high school level prior to FAA employment did not
predict ATC success in either a positive or negative direction. Essentially all of the controllers in the samples included in this study had
at least a high school education.

Other experimental instruments -- 16 Personality Factor Questionnaire,
Concept Adjective, and Biographical Information Questionnaire -- did not
add appreciably to the predictive capability of the experimental test
battery, the PEQ, and the OKT.

The validities derived from the analysis of the combined predictors
against the Aggregate ATC Success Criterion, by ATC option, for all year
groups combined, are provided in Table 4. Weights for the experimental
tests were derived from the multiple regression analysis. Separate weights
were developed for each test for each ATC option and for all ATC options
combined. The derived weights for all ATC options combined were MCAT Conflict -- 37; MCAT Aptitude -- 21; Directional Headings -- 15; and, Dial
Reading -- 0.

Academy trainees (1976) in the Terminal option were not identified
by IFR or VFR options. Consequently, they were included in both the IFR
and VFR option Ns, but only counted once in the total N for all ATC options.

The report by Mies and Colmen (1976, Appendix A) presented the validity
coefficients derived for the experimental test ba-tery (unweighted), prior
experience, ATC Occupational Knowledge Test (OKT), and education level for
each of the four ATC success criteria (excluding the aggregate criterion)
by ATC option. Appendix B showed the validity of selected experimental
tests with the aggregate ATC success criterion.

367.

Table 4

Validities of combined predictors
against the aggregate ATC success criterion
(by ATC option, all years combined)

ATC Option	Weighted Test Battery Scores		Weighted Test Battery plus PEQ		Weighted Test Battery and PEQ plus OKT	
	N	r	df	R	df	R
FSS	196	.23**	193	.26**	192	.26**
VFR	479	.26**	474	.44**	423	.45**
IFR	499	.26**	494	.39**	443	.43**
ARTCC	445	.30**	425	.32**	388	.37**
All Options	1309	.26**	1287	.32**	1205	.34**

** = $p \leq .01$

Placement. The weighted experimental test battery did differentiate between the FSS, Terminal (IFR, VFR) and ARTCC options. Average scores for ATC Specialists in FSS terminals and ARTCCs were different at statistically significant levels of confidence. The average FSS score was lowest and ARTCC highest. Table 5 provides the comparative mean scores for the various ATC options and year groups.

By analysis of variance it was found that the score means differed significantly (at the 1% level of confidence) between FSS, Terminal (VFR and IFR) and ARTCC for the 1969-70 and 1973-74 groups. The 1976 ATC group difference between terminal and ARTCC was significant at the 5% level

Test Fairness. Women as a group scored lower on each of the predictors and on the aggregate criterion of ATC success. In each case, these differences were significant at the 1% level of confidence (Table 6).

Validities of the predictors against the aggregate ATC success criterion are provided in Table 7 for men and women. Except for prior experience, these validities were significant at the 1% or 5% levels. It should be noted that few of the 229 women in the prior experience sample had aviation-related experience. Consequently, these validity results were affected by lack of variance.

Comparable statistical analyses were made between minority and non-minority groups and between non-minority and blacks, as defined by OMB Circular A-46, Standards and Guidelines for Federal Statistics, (1974). The results with respect to the test battery, prior experience, and the ATC Occupational Knowledge Test, are shown in Table 8, 9, and 10, respectively.

Minorities as a group, and blacks as an element of the minority group, scored lower on each of the predictors and on the aggregate criterion. In each case, these differences were significant at the 1% level of confidence. The validities of the predictors with the aggregate criterion, for minorities and non-minorities, are provided in Table 11.

When blacks were analyzed separately from all other minorities, the validity of prior experience and the OKT was sustained (r = .259 and .256, both at the 1% level of confidence). The validity of the weighted test battery was .20, which did not reach the 5% level of confidence (p = .75). However, the difference in the validity for non-minorities and blacks was not statistically significant.

Conclusions

The results of this study supported the conclusion that several of the experimental predictors were sufficiently valid and fair in predicting ATC success that they should be further developed for operational use in the selection of applicants for the air traffic control occupations. However, the lack of CSC test scores for a sufficient number of the sample

Table 5

Mean weighted test battery scores
by ATC option

ATC	1969-70 & 1973-74 ATC Hires			1976 New ATC Hires (Academy Trainees)		
Option	N	Mean	SD	N	Mean	SD
FSS	196	227.4	40.2	--	--	--
VFR (Terminal)	170	244.6	35.8)			
IFR (Terminal)	189	250.0	37.3) ------	310	242.0	35.0
ARTCC	182	264.3	36.6	263	247.9	38.4

Table 6

Means, standard deviations for men and women
and t-test results for predictors and the
aggregate ATC success criterion.

	Predictors									Aggregate Criterion		
	Weighted Test Battery			PEQ			OKT					
Sex	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Men	1397	244.5	380	2736	3.7	2.9	1318	76.6	12.3	1254	3.1	1.4
Women	171	234.5	480	235	.8	1.6	158	64.6	16.3	158	2.6	1.4
	(t=3.14; $p \leq .01$)			(t=15.02; $p \leq .01$)			(t=11.21; $p \leq .01$)			(t=4.56; $p \leq .01$)		

Table 7

Validities of the weighted test battery,
PEQ, and OKT against the aggregate ATC
success criterion, by sex groups.

Sex	Weighted Test Battery		Prior Experience (PEQ)		ATC Occupational Knowledge (OKT)	
	N	r	N	r	N	r
Men	1386	.23**	2721	.21**	1308	.22**
Women	165	.19**	229	-.04	154	.14*

*=p≤ .05
**=p≤ .01

Table 8

Test Sample

	Predictor Weighted Test Battery			Aggregate Criterion		
	N	Mean	SD	N	Mean	SD
Non-Minorities	1323	247.7	371	1308	3.2	1.4
All Minorities	245	220.4	429	243	2.7	1.4
	(t=10.30; p≤ .01)			(t=5.09; p≤ .01)		
Non-Minorities	1323	247.7	371	1308	3.2	1.4
Blacks	145	207.4	407	144	2.4	1.3
	(t=12.30; p≤ .01)			(t=6.31; p≤ .01)		

Table 9

Prior Experience Sample

	Predictor Prior Aviation Experience (PEQ)			Aggregate Criterion		
	N	Mean	SD	N	Mean	SD
Non-Minorities	2115	3.9	2.9	2097	3.2	1.4
All Minorities	321	2.4	3.0	318	2.6	1.4
	(t=8.48; p≤ .01)			(t=7.15; p≤ .01)		
Non-Minorities	2115	3.9	2.9	2097	3.2	1.4
Blacks	194	2.4	2.8	193	2.4	1.3
	(t=10.27; p≤ .01)			(t=8.39; p≤ .01)		

Table 10

ATC Occupational Knowledge Sample

	Predictor ATC Occupational Knowledge Test			Aggregate Criterion		
	N	Mean	SD	N	Mean	SD
Non-Minorities	1247	76.5	12.8	1235	3.2	1.4
All Minorities	229	69.1	14.7	227	2.7	1.4
	(t=7.85; p≤ .01)			(t=4.36; p≤ .01)		
Non-Minorities	1247	76.5	12.8	1235	3.2	1.4
Blacks	134	66.9	15.2	133	2.5	1.3
	(t=8.12; p≤ .01)			(t=8.12; p≤ .01)		

Table 11

Correlations for the weighted test battery, PEQ
scale, and OKT against the aggregate criterion
by minority/non-minority status.

	Weighted Test Battery		PEQ		OKT	
	N	r	N	r	N	r
Non-minority	1308	.203**	2097	.148**	1235	.201**
Minority	243	.219**	318	.247**	227	.272**

** $p \leq .01$

374.

popoulation, and the restriction of range on those scores which were available, made it impractical to compare the validity of the CSC test battery with the validity of the experimental tests. In order to address these questions, the Civil Service Commission, now the Office of Personnel Management (OPM), and FAA jointly directed efforts to botain the necessary information, for a group of about 11,500 applicants for ATC work during the Fall of 1976 and Spring of 1977 (See Chapter 19).

III

1976-1978 SELECTION STUDY OF NEW APPOINTEES TO THE ATC OCCUPATION

James O. Boone

Objective

This study was initiated by the FAA Civil Aeromedical Institute (CAMI) in July 1976 to evaluate the CSC Test Battery and new experimental predictors in relation to the success of newly hired ATC trainees in passing Initial Qualification training for the EnRoute and Terminal ATC options at the FAA Academy. Upon entry on duty, new ATC trainees are given a 2-week general orientation at their ATC facility (or regional headquarters). They then enter Initial Qualification training at the FAA Academy in Oklahoma City for the ATC option to which assigned. On their first day at the Academy, they are given a series of experimental tests and questionnaires. Their participation in the test program is voluntary. The experience has been that most of the students volunteered. Extensive efforts were made to obtain the individuals' test scores on the five-part *CSC Battery and their total earned ratings*, which include additional credit for veterans preference and aviation-related experience as provided by by the CSC Rating Guide. This information was correlated with their subsequent training status to evaluate the CSC Battery and to identify the experimental tests that were most predictive of training success.

Sample Description

During the period July 1976 through April 1978, 3,008 students entered Terminal and EnRoute ATC training at the FAA Academy. Selected characteristics of this total group are provided in Table 12.

The analysis of the CSC and experimental predictors required the use of complete information on the students. The 138 students who withdrew did not have laboratory criterion scores and were eliminated from the sample. The students for whom prior experience information was missing were also excluded. And an additional 824 students were excluded because they did not have one or more test scores (either the CSC test battery, taken prior to employment with FAA, or the experimental battery, because they did not volunteer to take the experimental tests).

375.

Table 12

Description of the 1976-1978 total trainee sample,
by pass-fail in training, sex, and aviation-related
experience.

| | | Pass | | Fail | | | |
| | Total N | N | Percent | Fail | | Withdrew | |
				N	Percent	N	Percent
Men	2580	2034	79	436	17	110	4
Women	428	272	64	128	30	28	6
Total	3008	2306	77	564	19	138	4
No Experience	834	579	69	206	24	58	7
Aviation-Related Exp.	1936	1579	82	296	15	61	3
Unknown	229	148	65	62	27	19	8
Total	3008	2306	77	564	19	138	4

The final sample included in the analysis, after these exclusions, consisted of 1,827 students distributed as shown in Table 13. The failure rate for this sample (19.2%) was essentially the same as for the total population (19.7%), after withdrawals were excluded.

Predictors

The following predictor tests were included in this study:

Civil Service Test Battery

   (CSC 24, Arithmetic Reasoning; CSC 51, Spatial Relations; CSC 135, Following Oral Directions; CSC 157, Abstract Reasoning; and CSC 540, Air Traffic Problems

Directional Headings Test (DHT) Total Score, Parts A and B.

   This was the same version as used in the 1977 EPA study (Mies, Colmen, and Domenech, 1977) and also in the 1976-1977 applicant study (Chapter 19).

Multiplex Controller Aptitude Test (MCAT) Total Score
   (Aptitude - A, Conflicts - C)

   Several different versions were used (606A, 606B, 706A, 706B, 607, 707). These varied in the number and mix of questions (aptitude and conflict) and in the length of time allowed for the test.

Dial Reading (DL RD)

   This was the same as in the 1977 EPA study.

ATC Occupational Knowledge Test (OKT) Form 101B.

   This was essentially the same as the OKT used in the 1977 EPA study.

Table 14 provides the means and standard deviations for the sample of 1,827 ATC trainees at the FAA Academy on the various predictors and compares them to results obtained with the 1976-1977 ATC applicant group, as well as with the ATC sample group included in the 1977 EPA study (Mies, Colmen, and Domenech, 1977), where comparable statistics were available.

Criterion Measure

In this analysis, the ATC laboratory average score was used as the criterion measure of ATC training success. Selection of this criterion was based on several considerations. First, prior studies had demonstrated that these scores were the most reliable predictors of subsequent success as an Air Traffic Controller. Second, the laboratory training phase consists of a series of operational air traffic control problems in which

377.

Table 13

Description of the 1976-1978 final trainee sample,
by pass-fail in training, sex, and aviation-related
experience

	Total N	Pass		Fail		Percent of Original Sample
		N	Percent	N	Percent	
Men	1587	1314	83	273	17	61.5
Women	240	162	68	78	32	56.1
Total	1827	1476	81	351	19	60.7
No Experience	558	409	73	149	27	66.9
Aviation-Related Exp.	1269	1067	84	202	16	66.5
Total	1827	1476	81	351	19	65.7

Table 14

Means and standard deviations on predictor tests
for 1976-1977 applicant sample[1], 1977 EPA study
sample[2], and present study sample[3].

Test	1976-1977 ATC Applicants				1977 EPA Study (N=1323-1229)		1976-78 ATC Trainees (N-1827)	
	CSC TOTAL (N=7412-6821)		CSC PASS (N=3690-3340)					
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CSC-24	39.66	9.6	44.80	6.8	NA		47.07	6.7
CSC-51	26.65	6.7	30.74	3.8	NA		31.90	3.3
CSC-540	28.98	13.1	37.70	10.1	NA		42.88	9.7
CSC-157	29.29	10.3	36.65	6.0	NA		38.17	6.4
CSC-135	21.50	8.6	27.00	5.1	NA		29.31	4.2
MCAT (A)	16.59	5.7	20.27	4.1	NC		23.14	4.5
MCAT (C)	9.17	4.2	11.38	3.9	NC		15.74	4.1
MCAT TOT	22.76	9.1	31.65	7.2	NC		38.87	7.6
DHT A	24.20	12.0	30.69	9.8	NA		32.68	9.0
DHT B	22.80	11.7	29.25	8.9	31.9	6.9	31.63	9.0
DHT TOT	47.00	22.6	59.93	17.3	NA		64.34	17.0
DL·RD	NA		NA		39.7	9.4	40.97	9.2
OKT	NA		NA		74.9	13.6	66.96	16.3

[1] See Chapter 19.

[2] Mies, Colmen, and Domenech, 1977

[3] Present study

NA - Statistics not available since the tests were either not administered
to these groups or could not be obtained from available records.
In the case of DHT, for the 1977 EPA Study, only Part B was scored
and used in the analysis.

NC - Statistics not comparable. In the 1977 EPA Study, two forms of MCAT
were administered to the test sample (706 A, 606 A). In the analysis,
all aptitude questions from both forms were combined for a total
aptitude score (mean = 42.8; SD = 6.2), as were all conflict questions
(mean = 29.4; SD = 5.6).

students must demonstrate their ability to apply the academic knowledge and skills acquired in training. Finally, this phase is conducted on a pass-fail basis.

The laboratory score constitutes 65% of the students' total score in the laboratory phase of training. Almost all of the students who fail Initial Qualifications training do so in the laboratory training phase. During the period from July 1976 through April 1978, several changes were made in the laboratory training phase which affected the pass-fail ratio of students. In July 1976, the weight given to the laboratory average was increased from 35% to 65%. In September 1976, the number of graded laboratory problems was increased from four to six. And in May 1977, "procedural errors" were incorporated in the grading for failure to handle aircraft in a timely manner during the problem exercise. In order to accommodate these changes and the differences between the EnRoute and Terminal laboratory procedures, the laboratory averages were converted to standard scores and data from the two ATC options combined. The standardized laboratory average criterion was labeled "ZLab." The reliability of the laboratory average for EnRoute training is .80 (2,223 students) and for the Terminal option .81 (1,982 students).

A detailed discussion of the ATC Initial Qualifications training program and follow-on training at the ATC facility level for the Terminal and EnRoute options is included in Chapter 9.

Analytical Methodology

Several different experimental forms of the MCAT were employed in the testing at the FAA Academy and the order of administration varied for each form. Consequently, it was necessary to standardize the scores for these tests. Since MCAT 706A was also administered to 6,821 ATC applicants by OPM in 1976-1977, the scores on the various forms of MCAT used by CAMI were standardized by linear conversions using the same metric as MCAT 706A. The order effect was handled by using the scores from MCAT 706A given first. The methodology used for these conversions is described in Dailey and Pickrel (1977) and Boone (1979).

A major problem encountered in the evaluation of the CSC and experimental tests was the restriction of range effect (See Thorndike, 1945), occasioned by the fact that the criterion (ZLab) information was available only on those individuals who were hired. This resulted in a spuriously low correlation between the CSC test scores and the criterion. In order to adjust the restricted correlations so that they would reflect the true relationship between the tests and the criterion for the applicant group, the correlations were corrected for restriction of range. The usual methods for this correction in the three-variable case assume that unrestricted information is available only on the variable used for selection, or the third incidental variable, but not both. However, with the information derived from the 1976-1977 ATC applicant population (Chapter 19), unrestricted data were available on both variables. A modified procedure was therefore developed to make use of all of the available information in the correction for range restriction. A full discussion of the procedure and the derivation of equations is provided in Boone (1978).

380.

The unrestricted correlation matrix from the total 1976-1977 ATC applicant sample (Chapter 19) that was used to correct for range restriction is provided in Table 15. The means, standard deviations and sample sizes (N's), except for the Earned Rating, are shown in Table 14. The Earned Rating for the applicant group was derived by taking the raw weighted CSC total score, transmuting it to a metric scale of 0-100 and, where appropriate, adding veterans preference (5 points) and additional credit for experience (0, 5, 10, or 15 points based on the CSC Rating Guide criteria) to the scores of applicants who passed the CSC Test Battery. This Earned Rating is the basis on which OPM ranks and places eligible condidates on a register form from which selections are made in accordance with OPM regulations.

Table 16 provides the restricted correlation matrix for the present sample of 1,827 ATC trainees at the ATC Academy. The means and standard deviations for this sample are also shown in Table 14. It should be noted that standardized (Z) scores (mean = 0; SD = 1) were used for the MCAT in computation of the intercorrelations.

The correlations of particular interest are those between the tests and the ZLab criterion. These are zero order validity coefficients. The effects of restriction are immediately apparent in the low correlations between the CSC tests, which were used in the actual selection, and the ZLab criterion. The two highest zero order coefficients were MCAT TOTAL (.277) and D1 RD (.272), neither of which was restricted directly by selection of the trainees. Two CSC tests, 51 -- Spatial Patterns and 135 -- Following Oral Directions, were omitted from the matrix. They were eliminated based on the negative skew found in other studies. Extreme selection, as in the case of ATC trainees, results in a sharp reduction of the variance in the selected groups; this was shown in Table 14, which compared means and standard deviations for applicant, applicant pass, and trainee groups.

In the correction for restriction of range, the difference between the applicant group variance and the selected group variance is employed as a measure of the amount of curtailment that occurs due to selection. It was not determined whether the skew resulted in a violation of the linearity assumption; however, the extreme disparity between the two variances for CSC 51 and CSC 135 resulted in a corrected correlation that was much higher than the other corrected correlations. In this analysis, if the correlations for CSC 51 and CSC 135 had been included in the multiple regression analysis, none of the other tests, either independently or in combination, would have added anything significant to the multiple R beyond that contributed by CSC 51 and CSC 135. These results were considered spurious and CSC 51 and CSC 135 were excluded from the multiple regression analyses reported below.

Table 17 presents the estimated unrestricted correlations (as well as the actual unrestricted correlation for the CSC tests from Chapter 19, Table 6). The correlations of primary interest are the correlations of the tests with the ZLab criterion. After correcting for range restriction,

381.

Table 15

Intercorrelations (unrestricted) of predictor tests.
1976-1977 applicants (Chapter 19). N = 7500.
(Decimal points omitted).

Variable

Variable	CSC 24	CSC 51	CSC 540	CSC 157	CSC 135	MCAT A	MCAT C	MCAT TOTAL	DH.A	DH.B	DH. TOTAL	EARNED RATING
CSC 24	1.00	.33	.56	.50	.51	.52	.44	.53	.46	.47	.49	.66
CSC 51		1.00	.46	.59	.54	.60	.42	.57	.54	.54	.57	.75
CSC 540			1.00	.59	.58	.63	.56	.65	.55	.56	.59	.78
CSC 157				1.00	.63	.63	.49	.62	.53	.56	.57	.85
CSC 135					1.00	.66	.52	.65	.54	.58	.59	.77
MCAT A						1.00	.69	.94	.65	.67	.69	.77
MCAT C							1.00	.89	.54	.55	.57	.60
MCAT TOTAL								1.00	.66	.67	.70	.76
DH.A									1.00	.82	.95	.68
DH.B										1.00	.95	.69
DH.TOTAL											1.00	.72
EARNED RATING												1.00

382.

Table 16

Intercorrelations (restricted) of predictor tests and ZLab for present sample of 1827 trainees in 1976-1978. MCAT data were based on standard scores (M = 0, SD = 1). CSC 51 and 135 were not included.

Variable	CSC 24	CSC 540	CSC 157	DHT 1	DHT 2	DHT TOTAL	MCAT A	MCAT C	MCAT TOTAL	ZLAB	DL-RD
									Z Scores		
CSC 24	1.00	.33	.14	.25	.26	.27	.21	.15	.20	.10	.34
CSC 540		1.00	.15	.20	.22	.23	.20	.19	.22	.10	.29
CSC 157			1.00	.09	.14	.12	.24	.16	.23	.07	.28
DHT-1				1.00	.80	.95	.26	.27	.30	.21	.34
DHT-2					1.00	.95	.28	.27	.31	.22	.34
DHT TOTAL						1.00	.28	.28	.32	.23	.36
MCAT A							1.00	.60	.90	.25	.48
MCAT C								1.00	.88	.25	.45
MCAT TOTAL									1.00	.28	.52
ZLAB										1.00	.27
DL-RD											1.00

383.

MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

Table 17

Intercorrelations (unrestricted and corrected) of predictor tests and ZLab used in regression analyses. N = 1827 trainees in 1976-1978.

Variable	CSC 24	CSC 540	CSC 157	DHT-1	DHT 2	DHT TOTAL	MCAT A	MCAT C	MCAT TOTAL	ZLAB	DL-RD
									Z Scores		
CSC 24	1.00	.56	.50	.46	.47	.49	.52	.44	.53	.34	.52
CSC 540		1.00	.59	.55	.56	.59	.63	.56	.65	.39	.50
CSC 157			1.00	.53	.56	.57	.63	.49	.62	.40	.50
DHT-1				1.00	.82	.95	.65	.54	.66	.43	.50
DHT-2					1.00	.95	.67	.55	.67	.45	.51
DHT TOTAL						1.00	.69	.57	.70	.46	.52
MCAT A							1.00	.69	.94	.50	.61
MCAT C								1.00	.89	.48	.55
MCAT TOTAL									1.00	.53	.64
ZLAB										1.00	.47
DL-RD											1.00

384.

the MCAT at .531, Dial Reading at .466, and Directional Heading at .461 had the highest zero order validity coefficients.

Results

The unrestricted and corrected correlations were utilized in a series of step-wise multiple regression procedures. Each regression model was a refinement of the preceding model, with test scores regressed on ZLab. The first model excluded total scores for DHT and MCAT, since these are the sum of their part scores, which were included. The multiple correlation (R), the squared multiple correlation (R^2), and the beta weights for the predictors are shown in Table 18.

In succeeding models, CSC 540 was eliminated because it failed to contribute to prediction, and total scores were used for DHT and MCAT instead of part scores. The resulting multiple R was .567. Then Dial Reading was eliminated because of its marginal value in the 1972 and 1977 EPA studies, and DHT was eliminated because, in its current format (very brief and highly speeded) it was not suitable for operational use. The final results are presented in Table 19. The initial model yielded a multiple R of .569, and the final model, a multiple R of .541.

A factor analysis (principal axis analysis with varimax rotation) was performed to explore the characteristics of the test scores. The results are provided in Table 20. Two rather clear structures appear to underly these data. Factor 5 and Factor 1 were the largest factors, accounting for 42.99% and 22.72% of the variance, respectively. The three experimental tests (DHT T, MCAT T, and DL RD) and Zlab were highly loaded on Factor 5, while CSC 24 was very high on Factor 1, along with moderate loadings for CSC 540 and CSC 157. CSC 540 was highest on Factor 4 (but almost as high on Factor 5), and CSC 157 was high on Factor 3. Factors 2, 3, and 4 were singlets for the CSC tests.

The beta weights derived from the Final Regression Model (Table 18) were converted to raw score weights and then assigned unit weights. The following equation constitutes the composite score:

$$Y_C = 1 \ (CSC\ 24) + 2 \ (CSC\ 157) + 4 \ (MCAT)$$

where Y_C = the composite test score for the battery. Using the unit weights, the multiple R derived was .5354 compared to .5407 with beta weights; the corresponding R^2 values were .2867 with unit weights, and .2924 with beta weights.

A cross-validation study was performed to investigate the stability of the results. The sample was randomly separated into two groups, and the weights derived from the first sample were applied to the second sample to observe shrinkage in the multiple R. Distribution statistics and inter-correlations for the two groups are presented in Table 21. The regression results for the first sample are presented in Table 22.

385.

Table 18

Initial regression model.
CSC and experimental tests.

R = 0.569

R^2 = 0.324

Variable	Validity R w. ZLab	Beta Weight
CSC 24	.34	0.0071
CSC 540	.39	-0.0066
CSC 157	.40	0.0555
DHT 1	.43	0.0513
DHT 2	.45	0.0912
MCAT A	.46	0.1452
MCAT C	.50	0.1668
DL-RD	.64	0.1856

Table 19

Final Regression Model,
MCAT-T, CSC 24, and CSC 157.

R = 0.54
R^2 = 0.292

Variable	Validity R w. ZLab	Beta Weight
CSC 24	.34	.061
CSC 157	.40	.096
MCAT-T	.53	.439

Table 20

Principal Axis Analysis of Test Scores,
with Varimax rotation.

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
CSC 24 (Arith. Reas.)	.973	-0.055	-0.026	-0.018	0.258	-0.021
CSC 540 (A. T. Probs.)	.440	0.357	-0.113	-0.592	0.585	-0.056
CSC 157 (Abstr. Reas.)	.358	0.079	-0.730	0.008	0.588	-0.035
DHT T	.264	-0.107	0.024	-0.035	0.726	0.054
MCAT T	.272	-0.104	-0.007	-0.031	0.736	0.052
ZLAB	.283	-0.104	0.014	-0.042	0.767	0.056
DL-RD	.314	-0.089	-0.079	-0.070	0.774	0.032
(R^2) Pct. variance accounted for	22.72	2.53	7.91	5.14	42.99	0.21

Table 21

Cross-validation samples:
Distributions by race and sex.
Descriptive statistics, and
intercorrelations of 4 variables.

Sample 1

	Men	Women	Total
Black	47	17	64
Hispanic	14	3	17
Am. Indian	0	1	1
Asian	6	1	7
Eskimo	1	0	1
Other	730	94	824
Total	798	116	914

	Mean	SD
ZLab	0.028	1.007
CSC 24	46.998	6.871
CSC 157	38.490	6.538
MCAT	35.608	7.451

Correlations

ZLab	1.000	0.328	0.402	0.537
CSC 24		1.000	0.500	0.530
CSC 157			1.000	0.620
MCAT				1.000

Sample 2

	Men	Women	Total
Black	45	16	61
Hispanic	15	4	19
Am. Indian	0	0	0
Asian	7	1	8
Eskimo	1	1	2
Other	723	101	824
Total	791	123	914

	Mean	SD
ZLab	-0.020	0.990
CSC 24	47.026	6.853
CSC 157	38.252	6.244
MCAT	35.686	7.307

Correlations

ZLab	1.000	0.326	0.396	0.527
CSC 24		1.000	0.500	0.530
CSC 157			1.000	0.620
MCAT				1.000

Table 22

Regression results for the experimental test battery, Sample 1.

$$R = 0.545$$
$$R^2 = 0.297$$

Variable	Beta Weight
CSC 24	.035
CSC 157	.102
MCAT	.455

390.

Unit weights, derived from the beta values for Sample 1 were computed. These weights (CSC 24 = 1; CSC 157 = 2; and MCAT = 4) were then used to compute multiple R and R^2 for Sample 1 and Sample 2, with the following results:

Sample	Multiple R	R^2
1	.538	.290
2	.529	.280

Further discussion of the analysis and results for the CSC and experimental tests in this study can be found in Boone's (1979) report. The results supported the recommendation that CSC 24, CSC 157, and the MCAT be used in combination as a new selection battery for the screening of applicant air traffic controllers. Some information was provided, which indicated that a modified form of the Directional Headings test would be suitable for operational use and that a Dial Reading Test should be explored in future studies.

In the analyses discussed thus far, the ATC Occupational Knowledge Test (OKT) was excluded from the various regression analyses. The OKT is a "job-knowledge specific" test and was not intended for use as a qualifying examination to determine applicant eligibility for employment consideration. Rather, its use was intended to be limited to measuring ATC-related experience and knowledge, as a basis for granting extra point credit to those applicants who pass the competitive test battery. Consequently, the OKT was analyzed separately, as reported below.

SEPARATE STUDY OF THE OKT, 1976-1978 TRAINEE SAMPLE

Objective

This investigation was focused on the relationship of OKT to: (1) the then current method of granting extra credit, based solely on ATC related experience, (2) pass-fail status during Initial ATC Qualification training, and (3) an ATC selection test battery comprised of CSC tests 24 and 157, and MCAT, and (4) the combined estimated validity of this test battery and OKT.

Sample Description

This study was based on the same group of 1827 ATC trainees that was used for regression analyses of the CSC and experimental tests, discussed previously.

Predictor

The OKT (form 101B) is a 100-item multiple choice test, covering knowledge of air traffic control regulations, communications, flight service station work, navigation aids, weather, and radar. The score is the total

number of correct responses. The mean and standard deviation for the 1827 ATC trainees were 66.96 and 16.3, respectively. The Kuder-Richardson-20 internal consistency estimate of reliability yielded a reliability coefficient of .95 for a sample of ATC new hires.

Criterion.

The same criterion measure, ATC laboratory average used in the multiple regression study, was used in analyses of OKT.

Analytical Methodology

The 1827 ATC trainees were grouped into ATC-related experience categories, based on their responses to a biographical questionnaire administered on the first day of training. Four types of ATC related experience were identified.

1. Other experience (including air defense command, communications operator, and prior ATC training without ATC operational experience).

2. Pilot experience

3. VFR or non-radar ATC experience

4. IFR or radar control ATC experience

These four types of experience (together with "No ATC-related experience") were used to classify each of the 1827 trainees into 5 experience groups, as discussed in Chapter 16:

		N
Group 1:	No experience + Other experience only	595
Group 2:	Pilot or Pilot + Other experience	389
Group 3:	VFR + any additional experience except IFR	191
Group 4:	IFR + any additional experience except VFR	144
Group 5:	IFR + VFR + any additional experience	511

The OKT scores were correlated with experience and later correlated with the ZLab criterion. In addition, statistical analyses were made of each of the five experience groups, in relation to OKT score ranges and pass-fail status in ATC training.

Results

The correlation of OKT scores with experience was .61. The correlation of OKT with ZLab was .22, compared to .11 for experience credit. This indicates that while OKT was closely related to experience, it predicted success more accurately than the then existing method of awarding experience credit, based on the OPM Rating Guide.

The results of the statistical analysis of each of the five experience groups, their OKT scores, and ATC training pass-fail status, are presented in Table 23.

While the mean OKT score for the total 1827 ATC trainees was 66.96 (SD = 16.3), Table 23 shows marked differences among the five experience groups and, within each group by OKT score range as shown below.

Experience Group	N	Mean	SD	Percent of Failure
None + Other Experience	592	51.49	14.8	26.9
VFR + Any other (except IFR)	191	72.25	11.4	21.5
Pilot + Other Experience	389	69.94	10.9	15.2
VFR & IFR + Any other	511	78.00	9.7	14.3
IFR + Any other (except VFR)	144	76.28	9.5	13.2

Of the total 1827 ATC trainees, 870 (48%) scored below 70 on the OKT. Of these, 359 (19.6%) were in ATC-related experience groups which previously earned extra credit under the CSC Rating Guide, for this experience. The failure rate for the 359 was 26.2% as compared to 28.2% for the 511 trainees with no (or other) ATC-related experience, who also scored less than 70 on OKT.

As can be seen in Table 23, as OKT score ranges became higher, the proportion of failures generally continued to drop in each of the five experience groups. This suggested that rather than using the Rating Guide method to grant extra credit for experience, a method based on OKT scores would be significantly more predictive of ATC training success.

Up to this point, the OKT scores have been compared with the Rating Guide method of crediting ATC-related experience and both were compared in terms of prediction of pass-fail in ATC training. For OKT to be truly useful, it must contribute to the prediction of ATC training success, over and above which can be achieved by a new selection test battery.

It will be recalled that a test battery composed of CSC 24, CSC 157, and MCAT had a multiple R of .54 with ZLab, which accounted for 29% of the criterion variance. If the correlations of OKT with ZLab and the other test battery scores were corrected for restriction of range, using Thorndike's formula 7 (Thorndike, 1945), then the multiple regression coefficient, including OKT, could be estimated. The new multiple R, calculated to include OKT as well as the three tests in the battery, would be .60 and R^2 would be .36, indicating that 36% of the ZLab variance would be accounted for. This increase is significant ($F = 204.3$ $p < .001$), indicating that the addition of extra points using OKT scores would significantly improve prediction and supports the recommendation that OKT should be used as a basis for determining extra credit. The analysis of OKT summarized here is discussed in more depth in Chapter 16.

Table 23

Failure rates by OKT score ranges and OKT
means by experience groups.

OKT Score	Group 1 Other		Group 2 Pilot		Group 3 VFR		Group 4 IFR		Group 5 VFR + IFR		Total	
	Total	Pct. Fail	Total	Pct. Fail	Total	Pct. Fail	Total	Pct. Fail	Total	Pct. Fail	Total	Pct. Fail
0-59	430	29.3	67	26.9	25	32.0	11	27.3	23	43.5	556	29.7
60-64	37	27.0	47	23.4	19	42.1	9	33.3	23	34.8	135	29.6
65-69	44	25.0	55	9.1	21	28.6	13	23.1	46	23.9	179	20.1
70-74	34	20.6	77	13.0	36	19.4	18	16.7	61	16.4	226	16.4
75-79	22	9.1	67	10.4	37	13.5	34	11.8	117	14.5	277	12.6
80+	25	12.0	75	10.5	53	13.2	59	5.1	241	7.1	454	8.4
Total	592	26.9	389	15.2	191	21.5	144	13.2	511	14.3	1827	19.2
OKT Mean	51.49		69.94		72.25		76.28		78.0		66.96	
SD	14.8		10.9		11.4		9.5		9.7		16.3	

394.

REFERENCE NOTES

1. Federal Aviation Administration. Air traffic controller selection and retention task force report, December 1974.

2. Federal Aviation Administration. Air traffic controller selection and retention cost analysis, March 1975.

395.

Chapter 19

STUDY OF ATC JOB APPLICANTS 1976-1977[1]

Donald B. Rock, John T. Dailey, Herbert Ozur,
James O. Boone, and Evan W. Pickrel

Objectives

Historically, very little normative information has been available
on the characteristics of individuals who apply for positions as Air Traf-
fic Controllers. Consequently, this study was developed to collect and
analyze data on ATC applicants with respect to sex, education, prior
aviation-related experience, and the relationship between these variables
and scores on the CSC test battery and recently developed experimental
tests. One direct use of this information was to establish a basis for
comparing the validities of the CSC tests and the experimental tests.
This was accomplished by correcting for the restriction of range on the
CSC test scores for approximately 2,000 newly hired ATC trainees who at-
tended the FAA Academy during 1976-1977 and for whom CSC test scores were
available, as discussed in Chapter 18, Part III.

Sample Description

During the period November 1976 through January 1977, the Civil Serv-
ice Commission opened the competitive ATC examination in the FAA Eastern
and Southern regions. These included the areas encompassed by Regions II,
III, and IV of the Standard Federal Regions. Approximately 11,500 appli-
cants took the ATC Civil Service test battery during that period. Of this
group, about 7,500 also completed the experimental tests and the Prior Ex-
perience Questionnaire (PEQ). While the sample from the 11,500 population
could not be controlled, differences between the means, standard deviations,
and other statistics on the CSC tests for the 7,500 who completed the ex-
perimental tests and provided prior experience information on the PEQ, and
the 4,000 who did not, were computed and were not statistically significant.

Predictors

In addition to the five-part CSC test battery, two experimental tests,
MCAT 706A and the Directional Headings Test (DHT), were administered to the
applicant group. These test forms were the same versions as used in the
1977 EPA study, Chapter 18, Part II. The same form of the PEQ was also given.

All tests were administered by CSC examiners as part of the normal com-
petitive testing procedures. Applicants were informed that the experimental
tests would have no bearing on their eligibility status and the experimental
predictors were administered after completion of the competitive CSC test
battery. The DHT and PEQ were scored manually and converted to computer

---

[1]Prepared by S. B. Sells

records by the FAA under contract. Determinations on applicants' sex were coded on the PEQ, based on the applicant's name. At the time, Federal regulations prohibited obtaining any ethnic or minority group information from the applicants.

Criteria

Since this was a normative study, limited to ATC applicants, it was not feasible to establish operational criterion measures of test validity. However, statistical analyses were made based on men and women applicants, and those who passed the CSC test battery, in contrast to those who failed.

Analytical Methodology

The data obtained on this ATC applicant group were analyzed with respect to frequency distributions, mean, and standard deviations, and by analysis of variance and intercorrelation of predictors.

Results

Test results. Table 1 shows the means and standard deviations for the total applicant group, the pass group (Those who scored above 209 on the CSC battery), and the fail group, for men and women applicants. It is evidence that over all, women scored somewhat below men on most of the CSC tests and the two experimental tests. These differences, while statistically significant due to the large sample size, were not important. For example, further analysis showed that sex accounted for only 1% of the variance in the CSC total test score, 3% in the MCAT, and 2% in the Directional Headings total test scores. In other words, most of the variance in these three predictors was related to factors other than sex.

In the pass group, the mean CSC test score for men and women was essentially the same. Consequently, discounting veterans preference and additional credit allowed for aviation-related experience, women who passed the CSC test should have had about the same opportunity to be selected for the ATC occupation as men.

The frequency distributions for each of the five CSC tests and the two experimental tests were also examined. Two of the CSC tests, CSC 135 and CSC 51, showed a marked negative skew (-1.30 and -1.80, respectively) and consequently, provided little differentiation between the applicants. This is shown in Figure 1.

Education and aviation-related experience in relation to the CSC battery. Table 2 provides descriptive information on education and experience levels for men and women in the applicant group. Aviation-related experience data were obtained from the Prior Experience Questionnaire (PEQ). This identified the specific work or skill elements which, based on the OPM Rating Guide, allowed extra credit points to applicants who passed the CSC test battery. About 22% of the total applicants were women. However, the distribution of

398.

Table 1

Means and standard deviations of men and women who passed and failed the CSC Battery, on the CSC Tests and the experimental MCAT and Directional Headings Test.

CSC Tests

Test Sex	Total Group N	Mean	SD	Pass Group N	Mean	SD	Fail Group N	Mean	SD
CSC-24									
Total	7412	39.66	9.6	3960	44.80	6.8	3722	34.61	9.3
Men	5720	39.63	9.5	2980	44.51**	6.9	2740	34.32**	9.2
Women	1607	39.83	9.7	663	45.91	6.6	944	35.56	9.3
CSC-51									
Total	7412	26.65	6.9	3960	30.74	3.8	3722	22.60	6.4
Men	5720	27.40	6.4	2980	31.05	3.6	2740	23.43	6.4
Women	1607	23.98**	7.0	663	29.37**	4.4	944	20.20**	6.0
CSC-135*									
Total	7412	21.49	8.6	3960	27.00	5.1	3722	16.03	7.8
Men	5720	21.88	8.4	2980	26.97	5.1	2740	16.35	7.8
Women	1607	20.13**	9.0	663	27.11	5.0	944	15.23**	7.9
CSC-157*									
Total	7412	29.29	10.3	3960	36.65	6.0	3722	21.99	8.3
Men	5720	29.47	10.2	2980	36.40**	6.0	2740	21.93	8.2
Women	1607	28.62**	10.9	663	37.75	5.8	944	22.21	8.0
CSC-540									
Total	7412	28.98	13.1	3960	37.70	10.1	3722	20.34	9.6
Men	5720	29.78	12.8	2980	37.84	9.9	2740	21.03	9.5
Women	1607	26.11	13.7	663	37.00	10.8	944	18.46**	9.8
CSC TOTAL									
Total	7412	202.03	52.3	3960	244.26	23.2	3722	160.16	37.3
Men	5720	205.04	50.9	2980	244.23	22.9	2740	162.42	36.8
Women	1607	191.29**	55.2	663	244.26	24.0	944	154.08**	37.8

Experimental Tests

Test Sex	Total Group N	Mean	SD	Pass Group N	Mean	SD	Fail Group N	Mean	SD
MCAT									
Total	6822	25.76	9.1	3340	31.65	7.2	3481	20.12	6.9
Men	5241	26.57	9.0	2688	32.14	7.0	2552	20.72	7.0
Women	1498	22.95**	8.7	607	29.53**	7.2	891	18.46**	6.5
DIR HEADING									
Total	7073	46.99	22.6	3583	59.93	17.3	3489	33.68	19.4
Men	5462	48.87	21.9	2899	60.63	16.6	2562	35.56	19.5
Women	1526	40.17**	23.3	637	56.58**	19.4	889	28.41**	18.2

*CSC 135 (Following Oral Directions) and CSC 157 (Abstract Reasoning) tests were double-weighted in computing the CSC test total score.

**Significant at the 1 percent level of confidence.

CSC – FOLLOWING ORAL DIRECTIONS
135   (DOUBLE WEIGHTED)

CSC – SPATIAL RELATIONS
51

MEAN = 21.5
S.D. = 8.6

MEAN = 26.7
S.D. = 6.7
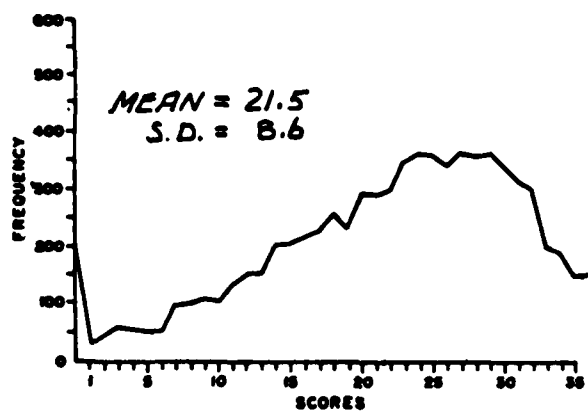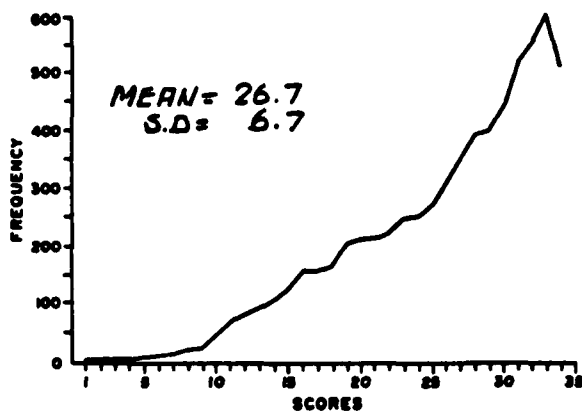
Figure 1.   Frequency distributions for CSC 135 and CSC 51, showing
            negative skew.

Table 2

Distribution of men and women applicants by
education and aviation-related experience.

Education Level	Total Applicants	
	N	Percent
Men with:		
High school or less	1225	21
Less than 3 years college	2352	41
3 or more years college	2126	37
Total	5703	100
Women with:		
High school or less	355	22
Less than 3 years college	618	39
3 or more years college	630	39
Total	1603	100
Experience Level		
Men with:		
No aviation-related experience	4393	77
Aviation-related experience	1327	23
Total	5720	100
Women with:		
No aviation-related experience	1483	92
Aviation-related experience	124	8
Total	1607	100

Education and Experience	Percent	
	Men	Women
With aviation-related experience		
High school or less	21	23
Less than 3 years college	44	39
3 or more years college	35	38
No aviation-related experience		
High school or less	22	22
Less than 3 years college	40	39
3 or more years college	38	39

401.

education level for men and women was approximately the same: 78% of both the men and women had some education beyond high school. With respect to aviation-related experience, 23 percent of the men and only 8 percent of the women identified experience which could result in the granting of extra credit points for selection eligibility. When education and experience were combined, the distribution of education level remained relatively consistent for both men and women and for those with and without aviation-related experience.

From this information it is evident that the level of education for men and women was essentially the same and did not differentiate between the applicants based on their sex. The distribution of aviation-related experience between men and women did differentiate between the applicants; 23% of the men were potentially eligible for extra credit, in contrast to only 8% of the women.

Pass-fail on the CSC battery. The applicant sample was analyzed in terms of pass-fail on the CSC test battery, in relation to sex, education, and experience, as shown in Table 3. In total, about 50% of the applicants passed the CSC battery. However, proportionally more men (52%) passed than women (41%). With respect to education levels, it is evident that those applicants with the most education had a significantly better chance of passing the test battery. Applicants with 3 or more years of college passed at about twice the rate of those who did not go beyond high school. A similar relationship also held for experience. Applicants with aviation-related experience also passed the CSC battery at about twice the rate of those without such experience.

Examination of pass-fail rates on the test battery in relation to education and experience combined disclosed that: (1) the level of education was distributed essentially in the same proportion between the experienced and non-experienced groups, and (2) experience increased the pass rate in each of the education groups. Applicants with no college or experience passed at a 27% rate. Those with 3 or more years of college and aviation-related experience passed at a 79% rate. Table 4 summarizes these data and shows the effects of education and experience on the percentages who passed the CSC battery.

Education and aviation-related experience in relation to individual CSC and experimental tests. Two-way analyses of variance were carried out for each CSC and experimental test. For this purpose, experience was classified further into sub-groups, since the CSC Rating Guide provided different levels of extra credit based on specific types of experience. The experience and education subgroups were as follows:

Aviation-Related Experience

1 No aviation-related experience
2 Communication experience only
3 Non-pilot with IFR or Air Defense Command (ADC) experience

402.

Table 3

Distribution of total applicants and those who passed
and failed the CSC battery, by sex, education, and
aviation-related experience.

	Total Applicants		CSC Battery Status	
	N	Percent	Pass Pct.	Fail Pct.
Men	5720	78	52	48
Women	1607	22	41	59
Total	7327	100	50*	50*
Education Level				
High school or less	1581	22	32	68
Less than 3 years college	2973	40	49	51
3 or more years college	2757	38	61	39
Total	7311	100	50*	50*
Experience Level				
No aviation-related experience	5876	80	45	55
Aviation-related experience	1451	20	68	32
Total	7327	100	50*	50*
Education and Experience				
Aviation-related experience, with				
High school or less	308	21	52	48
Less than 3 years college	632	44	67	33
3 or more years college	508	35	79	21
Total	1448	100	68*	32*
No Aviation experience, with				
High school or less	1273	22	27	73
Less than 3 years college	2341	40	44	56
3 or more years college	2222	38	57	43
Total	5836	100	45*	55*

*Pass-Fail rates for the Total Group

Table 4

Effects of education and aviation-related experience on the percentages of applicants who passed the CSC battery.

	CSC Test Battery Pass Rate		
Education Level	Percent with No Aviation-Related Experience	Percent with Aviation-Related Experience	Percent Increase
High school or less	27	52	+25
Less than 3 years college	44	67	+23
3 or more years college	57	79	+22

4   Non-pilot with VFR ATC experience
5   Pilot with IFR or ADC experience
6   Pilot experience only
7   Pilot with VFR ATC experience

Education

1   High school or less
2   Less than 3 years of college
3   3 or more years of college

Table 5 provides the results of the analysis variance for the CSC total test score. The lowest mean score (173.10) is for the no-experience and no-college group; the highest mean score (241.00) is for the pilot with VFR experience and 3 or more years of college group. The analysis of variance showed that experience and education were both highly significant sources of variance in the CSC battery scores, although the interaction of experience and education was not a significant factor. Furthermore, upon further analysis, it was found that neither, although related significantly, accounted for sufficient variance to be important; neither one accounted for more than 1% of the variance in the CSC total score.

However, when the CSC total score was categorized into pass and fail groups, education and experience together accounted for nearly 10% of the variance in pass-fail status. On this basis, the effects of education and experience are important. This same relationship was found to apply to the two experimental tests as well.

In summary, the data examined up to this point have demonstrated: (1) that applicants with aviation-related experience had a higher probability of passing the CSC battery (68%) compared to those without such experience (45%), and (2) that few women have had aviation-related experience. However, these two findings are not sufficient to explain the lower pass rate on the CSC battery for women (41%) compared to that for men (52%). Applying these probabilities to the experience and no-experience groups, it would be expected that 752 women would have passed, but only 663 did pass. For men, 2,879 men would be expected to have passed, but 2,980 actually did pass. Consequently, the fact that men had more aviation-related experience does not explain why women scored lower on most of the tests.

Intercorrelations. The intercorrelations among the tests, education, and experience are shown in Table 6. These correlations were used by Boone (1979a, see Chapter 18, Part III) to correct for restriction of range on test scores of trainees who attended the FAA Academy during 1976, 1977, and 1978.

Several stepwise multiple regression analyses were carried out using pass-fail on the CSC battery as the criterion. Table 7 shows the results for the CSC test battery, the order in which the individual tests entered the analysis and their contribution to the R^2, which is a measure of the amount of variance in the criterion accounted for by each test.

405.

Table 5

Analysis of variance of Total CSC battery score by education
and aviation-related experience groups.

A. Numbers of applicants with complete data on education, experience, and CSC battery.

Educ. Groups	Experience Groups							
	1	2	3	4	5	6	7	Total
1	1239	34	140	28	19	67	54	1581
2	2266	75	216	47	56	180	133	2973
3	2179	70	62	25	31	230	170	2757
Total	5684	179	418	100	106	477	347	7311

B. Mean CSC scores, by group.

Educ. Groups	1	2	3	4	5	6	7	Mean
1	173.10	179.53	210.69	201.96	198.68	191.07	211.56	179.48
2	197.40	204.79	218.55	233.23	220.81	221.52	225.19	202.83
3	209.11	221.59	224.96	222.72	246.10	236.17	241.00	214.43
Mean	196.59	206.56	216.94	221.85	224.13	224.31	230.34	202.16

C. Analysis of Variance

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	Prob.
Row	248,889.27	2	124,444.63	51.25	<0.001
Col	152,242.89	6	25,373.82	10.45	<0.001
Row X Col	44,320.17	12	3,693.35	1.52	0.109
Within Cell	17,701,966.50	7290	2,428.25		

406.

Table 6

Intercorrelations of CSC and experimental tests, education, and experience. Total applicant sample.[1]

Variable	1 CSC 24	2 CSC 51	3 CSC 540	4 CSC 157	5 CSC 135	6 CSC Total	7 CSC Total Wt	8 MCAT-C	9 MCAT-A	10 MCAT Total	11 Dir. Head A	12 Dir. Head B	13 Dir. Head Total	14 Educ. Level	15 PEO. AGG.	16 Educ. Group	17 Exper. Group	18 Exper. Group -x
1 CSC 24	100	33	56	50	51	75	69	44	52	53	46	47	49	27	6	26	10	11
2 CSC 51		100	46	59	54	69	75	42	60	57	54	54	57	11	18	12	23	19
3 CSC 540			100	59	58	85	80	56	63	65	55	56	59	16	13	16	14	5
4 CSC 157				100	63	83	89	49	63	62	53	56	57	22	13	22	17	14
5 CSC 135					100	81	79	52	66	55	54	58	59	20	15	21	16	7
6 CSC Total						100	99	62	77	77	66	68	71	24	16	25	19	14
7 CSC Total Wt							100	61	77	76	66	68	71	24	17	24	20	15
8 MCAT-C								100	69	89	54	55	57	14	17	14	18	5
9 MCAT-A									100	94	65	67	69	18	22	18	25	12
10 MCAT Total										100	66	67	70	18	22	20	24	10
11 Dir. Head A											100	82	95	20	26	20	33	26
12 Dir. Head B												100	95	20	22	20	27	21
13 Dir. Head Total													100	21	26	21	31	25
14 Educ. Level														100	-6	21	5	25
15 PEO. AGG.															100	95	67	-37
16 Educ. Group																100	-6	22
17 Exper. Group																	100	100
18 Exper. Group-X																		100

407.

[1] Decimal points not included.

Table 7

Stepwise multiple regression of the CSC tests
on pass-fail on the CSC battery.

Variable	r	R	R^2	Beta Weight
CSC 157	.711	.711	.505	.305
CSC 540	.662	.772	.597	.253
CSC 51	.608	.794	.630	.203
CSC 135	.639	.802	.644	.142
CSC 24	.529	.806	.649	.100

Table 8 shows the corresponding regression results for the two experimental tests; again the criterion was pass-fail on the CSC battery. The difference between the R^2 for these tests (45%) compared to the CSC tests (65%) indicates that they differ in some degree from the CSC tests in what they measure.

In the case of the experimental tests, MCAT-A (aptitude), which measures aptitudes similar to those measured by the CSC battery, entered the regression first and accounted for 40% of the total variance.

Table 8

Stepwise multiple regression of the experimental
tests on pass-fail on the CSC battery.

Variable	r	R	R^2	Beta Weight
MCAT-A (aptitude)	.631	.631	.399	.372
Dir. Head. Total	.582	.662	.439	.257
MCAT-C (conflict)	.522	.668	.446	.117

Chapter 20

RESEARCH ON THE EXPERIMENTAL TEST BATTERY FOR ATC APPLICANTS

STUDY OF JOB APPLICANTS, 1978[1]

Donald B. Rock, John T. Dailey, Jerbert Ozur,
James O. Boone, and Evan W. Pickrel

INTRODUCTION

This is the first of three chapters that address research focused on
the new experimental test battery for selection of ATC trainees, that was
adopted for operational use by the Office of Personnel Management (OPM) in
October 1981. The new battery was developed in an extensive program of
research described in Chapter 18, which culminated in an analytic study
at CAMI by Boone, in which the validities of scores on the then current
selection battery, referred to here as the CSC (Civil Service Commission)
battery, and of two experimental tests (the MCAT and the OKT), were com-
pared for a sample of 1827 new ATC trainees at the FAA Academy during 1976
to 1978. The experimental battery was composed of the MCAT, the Abstract
Reasoning Test (CSC-157), and the Arithmetic Reasoning Test (CSC-24), plus
the OKT as a basis for additional credit to reflect occupational knowledge
and experience. The research described here and in the following two chap-
ters was initiated in 1978 to investigate the potential of this battery
for operational selection of ATC applicants and to compare the new battery
with the CSC battery on the basis of prediction results for a representa-
tive sample of applicants, as well as various subgroups of the applicant
population.

This study involved 6000 applicants for employment as ATCS's and was
undertaken in order to compare the results of the new ATC battery with those
of the CSC battery with respect to the composition of the respective groups
that passed these tests and the effects on score distributions within each
group. Race-ethnic and sex data had not been available in the earlier stud-
ies, but a change in policy made them available in 1978. Arrangements were
made between the FAA and the OPM to administer the parallel forms of the
MCAT and a 60-item version of the OKT when the CSC battery was administered
to ATC job applicants during September through November, 1978.

RESEARCH APPROACH

Sample Description

The CSC battery, together with MCAT and OKT, were administered to a
total of 6,000 ATC job applicants. Some 5,331 applicants were scheduled
for examination through established OPM procedures. In addition, FAA had
arranged with OPM for a series of "walk-in" test sessions which did not
require advance scheduling by applicants with OPM. A total of 669 appli-
cants took the tests on a walk-in basis. The walk-in sessions were

[1] Prepared by S. B. Sells

411.

established as a means of encouraging an increased number of women and minorities to compete for positions in the ATC occupation. Table 1 provides the distribution by race-ethnic group and sex for 5295 of the 5331 scheduled ATC job applicants. Table 2 presents the same data for 664 of the 669 walk-in applicants.

Several points of interest should be noted regarding the race-ethnic and sex distribution of these two applicant groups. For the scheduled group, 92% were black or white men and women, and 71% were white men and women. Generally men .epresented 70% - 79% of every race group except blacks; only 58% of the blacks were men, and 42% were women, almost twice the ratio of other groups.

The distribution of the walk-in applicants (Table 2) indicates that the objective of having more minorities and women compete for the ATC jobs was fairly successful. Some 47% of this group were women, in contrast to 28% of the scheduled applicants, and 55% were minorities, in comparison to 29% of the scheduled applicants.

Predictors

The tests included in this study consisted of the then existing CSC selection test battery and parallel forms of MCAT and OKT:

CSC Test Battery

CSC-24	Arithmetic Reasoning
CSC-51	Spatial Relations
CSC-135	Following Oral Directions
CSC-157	Abstract Reasoning
CSC-540	Air Traffic Problems

MCAT	OPM ID	ODD/EVEN TEST COMBINATION
Form	120	(4o6e)
Form	130	(4e6o)
Form	140	(6o7e)
Form	150	(6e7o)
Form	160	(7o4e)
Form	170	(7e4o)

OKT - 101c   (60-item version)

As in previous experimental testing of applicant groups, the MCAT and OKT were administered after completion of the regular CSC test battery. In addition to the tests, applicants completed forms provided by OPM on a voluntary basis to obtain race, sex, and other information for separate analysis by OPM.

Criterion and Analytic Methods

Since this study encompassed ATC applicants, it was not feasible to establish operational criterion measures of validity. However, a number of statistical analyses were made based on pass-fail eligibility comparisons, score distributions, and means and standard deviations on the various tests.

Table 1

Description of the sample of regularly scheduled
ATC applicants tested during September to
November, 1978, by sex and race-ethnic group.
N = 5295 of 5331 applicants tested.

	Amer. Indian	Asian	Black	White	Hispanic	Other	TOTAL
Men	42 (72%)	38 (76%)	650 (58%)	2860 (76%)	213 (79%)	21 (70%)	3824 (72%)
Women	16 (28%)	12 (24%)	463 (42%)	915 (24%)	56 (21%)	9 (30%)	1471 (28%)
TOTAL	58 (1%)	50 (1%)	1113 (21%)	3775 (71%)	269 (5%)	30 (1%)	5295 (100%)

413.

Table 2

Description of the sample of "walk-in" applicants tested during
September to November, 1978, by sex and race-ethnic group.
N = 664 of 669 "walk-in" applicants tested.

	Amer. Indian	Asian	Black	White	Hispanic	Other	TOTAL
Men	2	3	145 (50%)	160 (55%)	44 (65%)	-	354 (53%)
Women	1	4	146 (50%)	132 (45%)	24 (24%)	3	310 (47%)
TOTAL	3	7	291 (44%)	292 (44%)	68 (10%)	3	664 (100%)

RESULTS

Descriptive statistics for the scheduled and walk-in groups were developed for each test. It should be pointed out that of the 260 scheduled applicants who identified themselves as Hispanic, all but 74 also identified themselves with another racial group. For example, 139 were identified as Hispanic-white, and in this analysis, they were included in the white group. The 104 applicants in the "other" category shown in Table 3 consisted of 74 Hispanic applicants and 30 of some other racial groups not identified. Descriptive statistics for the scheduled group of applicants are provided in Table 3 and for the walk-in group, in Table 4. (In Table 4, the 24 applicants in the "other" category consisted of 21 Hispanic and 3 otherwise not identified.)

In the scheduled applicant group, the mean scores for women were examined for both the black and white groups to assess their relationships to the scores of men in the same racial groups. Table 3 indicates that women scored somewhat higher than men in each racial group on CSC tests 24 and 157, but somewhat lower on CSC 51, 540 MCAT, and OKT; the greatest difference between men and women occurred on the OKT. The most marked difference between racial groups was between the black and white groups; the mean test scores for blacks were generally about one standard deviation below the mean for whites. This held for both men and women in each of the groups, and is completely in line with expectations arising from the results of other testing programs.

Relatively few differences between mean scores of the scheduled and walk-in groups were evident by racial groups. Within the black group, walk-in applicants scored about 2 points higher on CSC 24 and 3 points higher on CSC-157. The white group walk-in applicants scored almost 2 points lower on MCAT and about 3 points lower on OKT. These lower scores may be a result of the higher percentage of women in the white walk-in group (45%), compared to the scheduled group (24%).

For the applicant group tested in 1976-1977 (Chapter 19), two of the CSC Tests (51 - Spatial Relations and 235 - Following Oral Directions) showed marked negative skew. It was of interest to determine whether this characteristic would be replicated for these two tests in the 1978 applicant population. The frequency distribution of test scores for these two CSC tests in the present sample showed that again they provided little differentiation among the applicants.

Veterans preference for the total sample of scheduled and walk-in applicants was identified and the distribution of veterans credit is shown in Table 5, by sex and race-ethnic group, for 5846 of the 6000 applicants for whom these data were available.

Since veterans preference points are added to the test battery score for those who pass the test, it is evident that only 6% of the women could

Table 3

Means and Standard Deviations of CSC and Experimental tests, by sex and race-ethnic group, for scheduled applicants, September - November, 1978. N = 5331.

	American Indian N = 66		Asian N = 59		Black N = 1146		White N = 3914		Other N = 104		TOTAL N = 5331	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
CSC-24	40.1	9.2	41.9	7.9	35.4	9.6	42.4	7.8	36.8	8.6	40.7	8.9
Men	-	-	-	-	34.3	9.7	42.2	7.9	-	-	-	-
Women	-	-	-	-	37.0	9.2	43.2	7.6	-	-	-	-
CSC-51	28.6	4.8	28.3	5.8	23.1	6.5	29.1	5.2	27.0	6.5	27.7	6.1
Men	-	-	-	-	24.3	6.4	29.6	5.0	-	-	-	-
Women	-	-	-	-	21.3	6.2	27.7	5.6	-	-	-	-
CSC-135	21.6	9.0	19.9	8.2	16.6	8.1	24.7	7.0	20.1	8.5	22.7	8.1
Men	-	-	-	-	16.7	8.0	24.8	6.9	-	-	-	-
Women	-	-	-	-	16.5	8.3	24.2	7.4	-	-	-	-
CSC-157	28.0	8.2	31.0	10.9	22.1	9.3	30.9	8.9	26.1	9.0	28.8	9.7
Men	-	-	-	-	21.5	9.3	30.5	8.9	-	-	-	-
Women	-	-	-	-	22.9	9.1	32.1	8.8	-	-	-	-
CSC-540	29.4	12.7	30.5	12.6	22.7	11.9	33.1	11.7	25.6	12.5	30.6	12.5
Men	-	-	-	-	23.4	11.9	33.6	11.6	-	-	-	-
Women	-	-	-	-	21.8	11.9	31.3	11.9	-	-	-	-
MCAT	31.1	8.6	31.0	8.6	22.1	7.4	34.2	8.1	28.0	8.4	31.4	9.5
Men	-	-	-	-	22.9	7.8	35.0	8.0	-	-	-	-
Women	-	-	-	-	22.1	6.8	31.8	8.2	-	-	-	-
OKT	24.3	10.8	24.9	11.3	17.4	8.7	27.2	11.7	21.8	11.8	24.9	11.8
Men	-	-	-	-	18.8	9.9	28.8	12.0	-	-	-	-
Women	-	-	-	-	15.5	6.5	22.0	8.9	-	-	-	-

416.

Table 4

Means and Standard Deviations of CSC and Experimental tests,
by sex and race-ethnic group, for walk-in applicants,
September-November, 1978. N = 669.

	American Indian N = 5		Asian N = 10		Black N = 297		White N = 328		Other N = 24		TOTAL N = 669	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
CSC-24	40.0	4.2	48.0	9.5	37.5	9.7	42.3	7.8	39.4	9.6	40.1	9.1
CSC-51	29.0	4.0	31.8	4.2	23.0	6.2	28.6	5.0	28.8	4.2	26.1	6.3
CSC-135	19.0	7.2	24.6	7.3	16.7	7.4	23.7	6.8	19.7	7.6	20.4	7.9
CSC-157	30.0	11.6	34.6	12.2	25.0	9.4	32.0	8.9	28.7	11.6	28.8	9.9
CSC-540	36.4	8.4	45.2	12.1	23.6	12.7	32.4	12.5	30.4	13.4	28.7	13.5
MCAT	28.4	3.9	34.9	7.4	22.7	7.5	32.6	7.8	27.9	9.7	28.0	9.1
OKT	19.2	6.8	16.4	5.5	16.2	6.5	23.7	10.9	19.5	8.9	20.1	9.7

417.

Table 5

Distribution of veterans credit by sex and race-ethnic group. Total sample of scheduled and walk-in applicants. N = 5846.

Veterans Credit Points	Men		Women		TOTAL	
None	2406	59%	1640	94%	4046	69%
5 Points	1571	38%	97 }	6%	1668	29%
10 + Points	123	3%	9		132	2%
TOTAL	4100	100%	1746	100%	5846	100%

Veterans Credit Points	American Indian		Asian		Black		White		Hispanic		Other		TOTAL	
None	42	70%	40	70%	981	73%	2724	68%	223	68%	27	84%	4037	69%
5 Points	17 }	30%	17 }	30%	318 }	27%	1210 }	32%	99 }	32%	5 }	16%	1666	29%
10 + Points	1		-		41		83		7		-		132	2%
TOTAL	60		57		1340		4017		329		32		5835	

418.

have benefited, and thus be ranked higher on the OPM register for appoint-
ment eligibility. This is in contrast to 41% of the men who could have
benefited from military service.

Eligibility for veterans preference points by race-ethnic groups was
about evenly distributed (30%-32%) except for the black group (27%), where
the higher ratio of women applicants to men reduced the number who would
have received veterans preference points.

Table 6 compares the total ATC applicant group to the civilian labor
force (CLF) by race-ethnic group, based on 1970 census data, with amend-
ments through 1977. Blacks as a group were significantly overrepresented
in comparison to their percentage in the civilian labor force, while whites
as a group were underrepresented. Women were also underrepresented; they
comprised 28% of the ATC applicant group, compared to 38% of the civilian
labor force.

Analysis of the CSC Battery

Scores of the total sample were analyzed in terms of the percentages
passing and failing the CSC battery and by sex and race-ethnic group.
These data are shown in Table 7. Because of the small numbers of Asian
and American Indian applicants, the analysis for minorities was confined
to the black and Hispanic groups. The minimum passing score on the test
battery was 70, which equated to the approximate mean score and was ap-
proximately equivalent to the 50th percentile.

Compared to the 7500 applicants tested in 1976-1977 (Chapter 19),
this sample had about the same pass rate for men (53% compared to 52% in
the earlier group) and a somewhat higher pass rate for women (45% compared
to 41%). In view of the large number of minorities who failed the CSC
battery in the present sample, analysis was made of the score distribution
by race-ethnic group. This is shown in Table 8. In this table, it can be
seen that the proportion of extremely low scores (50-59 and below 50) was
exceptionally high for the blacks (27% and 32%, compared to 17% and 18%
for Hispanics and 10% and 4% for whites).

As shown in Table 9, whites were underrepresented in the applicant
group, compared to their frequency in the civilian labor force, while
their frequency in the pass group was identical to their frequency in the
labor force. At the same time blacks were overrepresented among appli-
cants by a considerable amount and slightly underrepresented in the pass
group. Hispanics, American Indians, and Asians were represented in all
three groups, roughly in proportion to their frequency in the labor force.

Analysis of the OKT

Administration of the OKT to the 1978 applicant sample provided the
first opportunity to evaluate this test for a sizable sample of ATC appli-
cants. The OKT used was form 101C, a 60-item test. The data for this
test were reviewed for the total sample and separately for men and women

419.

Table 6

Comparison of the total applicant group with the civilian labor force[1]
by race-ethnic group.

	Civilian Labor Force	Percent of Applicants
American Indian	.3%	1%
Asian	.8%	1%
Black	9.7%	24%
White	85.0%	68%
Hispanic	4.2%	6%

[1]based on 1970 U.S. Census, with amendments through 1977.

Table 7

Distribution of total applicant sample by pass vs fail
on the CSC battery, by sex and race-ethnic group.
N = 5976.

Group	No. of Applicants	Pass		Fail	
		N	Percent	N	Percent
Men	4191	2236	53	1955	47
Women	1785	799	45	986	55
White	4067	2556	63	1511	37
Hispanic	339	128	38	211	62
Black	1407	264	19	1143	81
Other (not included)	(163)	(87)		(76)	
TOTAL	5976	3035	51	2941	49

Table 8

Distribution of CSC battery scores for the total sample,
showing pass and fail range separately, for white, Hispanic,
and black applicants. N = 5813.

Group	N	Pass Group			Fail Group		
		Over 89	80-89	70-79	60-69	50-59	Under 50
White	4067	453 (11%)	975 (24%)	1128 (28%)	920 (23%)	406 (10%)	185 (4%)
Hispanic	339	17 (5%)	41 (12%)	70 (21%)	92 (27%)	56 (17%)	63 (18%)
Black	1407	12 (1%)	63 (5%)	189 (13%)	305 (22%)	381 (27%)	457 (32%)
TOTAL	5813	482 (8%)	1079 (19%)	1387 (24%)	1317 (23%)	843 (15%)	705 (12%)

422.

Table 9

Frequency of race-ethnic groups in the civilian labor force,
the ATC applicant group, and the group that passed the
CSC battery. Total sample, September - November, 1978.

	Civ. Labor Force	ATC Applicants	Applicants in Pass Group
American Indian	.3%	1%	1.0%
Asian	.8%	1%	1.1%
Black	9.7%	24%	8.9%
White	85.0%	68%	84.8%
Hispanic	4.2%	6%	4.2%
	100.0%	100%	100.0%

and for race-ethnic groups, and the relation of test scores to veterans preference was investigated for the subsample of veterans.

Tables 3 and 4 (earlier) provided distributional data on the OKT for the various groups; Table 10 shows the means and standard deviations for scheduled and walk-in men and women. The walk-in applicants had a lower mean score (20.1) than the scheduled group (24.9), and this difference was evident among men and women and in all racial groups. The mean scores for blacks were 17.4 (scheduled) and 16.2 (walk-ins), which would indicate that there was relatively little difference in ATC-related knowledge among blacks in these groups, but that among other race-ethnic groups, the walk-in group had less ATC-related knowledge or experience than the scheduled group. In view of the objective of the walk-in recruitment program, these differences were expected.

In Table 11, the raw scores were converted to a scale of 0-100 and the Hispanic group was distributed among the other race-ethnic groups (e.g., those who claimed to be Hispanic-white were treated as white). Approximately 20% of the men scored above 64, compared to 4% of the women; and 19% of the whites, 12% of the Asians-American Indians, and 4% of the blacks, respectively, scored above 64.

Extra credit for experience. Previous studies with ATC trainees at the Academy and with developmental and journeyman ATC specialists had demonstrated a significant correlation between OKT scores and success in ATC training and work performance. These studies also supported the conclusion that the OKT provided a more effective basis for granting extra credit to ATC applicants who passed the CSC battery than the former Rating Guide that had been used for this purpose. Consequently, for further analysis, OKT scores of this applicant group were given the following extra points:

OKT Scores	Extra Credit
65-69	3 points
70-74	5 points
75-79	10 points
80+	15 points

These point values are similar to those provided by the OPM Rating Guide, which provided 5, 10, or 15 points, depending on the specific ATC-related experience.

In most cases, the types of ATC-related experience for which an applicant was granted extra credit were obtained through military service and training. For example, the Air Traffic Control experience which is granted the most points (10-15) was obtained through military experience. This also tended to be the case, although to a lesser extent, for pilot and other types of credited experience.

Table 10

Means and Standard Deviations of OKT scores of men and women
in the scheduled and walk-in applicant group. 1978 applicant
sample, N = 5976.

		Scheduled			Walk In	
Group	N	Mean	SD	N	Mean	SD
Men	3835	26.9	12.2	356	22.2	11.2
Women	1473	19.8	8.7	312	17.6	6.9

Table 11

Distribution of converted OKT scores (scale of 0-100) by sex and race-ethnic group.[1] 1978 applicants, N = 5976.

Percentage by OKT Score Range

Group	N	Below 15	15-24	25-34	35-44	45-54	55-64	65-74	75-79	Above 79
Men	4191	3	11	27	20	9	10	9	8	3
Women	1785	5	20	43	19	5	4	3	1	*
White	4242	2	9	27	22	10	10	9	7	3
Asian } American Indian	140	3	14	35	18	9	9	6	5	1
Black	1443	8	32	39	13	2	2	2	2	*

* Less than 1%.

[1] Hispanics distributed among other race-ethnic groups.

426.

In order to assess the relationship between military experience and OKT scores, an analysis was completed of applicants with veterans preference, their OKT scores, and the extent to which they could earn extra credit. The results are provided in Table 12.

Prior analysis of the relationship between OKT scores and pass-fail in ATC training for 1827 students (Boone, 1979a, See Chapter 18, Part III) indicated that the largest drop in total failure rate was between those trainees who scored 60-64 (29.6% failure rate) and those who scored 65-69 (20.1% failure rate). Consequently, for this analysis, comparisons were based on providing 3 points for those applicants who scored between 65-69. About 16% (968) of the total 6000 applicants could have been eligible for additional point credit based on OKT scores. A total of 702 (72%) of these also could have received points for veterans preference. The increasing number of applicants with veterans preference who scored "highest" on the OKT suggests that most of them probably had military air traffic control experience.

By contrast, those applicants who were not veterans constituted 28% of the total group who could receive extra credit based on OKT scores. While applicants in this group scored "lower" on OKT, use of this test to grant extra credit broadened the opportunity to receive extra credit for non-veterans who acquired ATC related knowledge through other experience.

Under the OPM rating and ranking procedures that were current in 1978, an applicant must have earned a passing score of at least 70 on the CSC test battery before additional credit could be given for veterans preference or aviation-related experience. Of the total 968 applicants who could have been eligible for extra credit based on OKT scores, 272 (28%) scored below 70 on the CSC test battery. Table 13 shows the distribution by point groups of these 968 applicants. The 696 who passed the CSC battery and would have been eligible for extra credit based on OKT scores represented 23% of the 3083 applicants who received passing scores on the battery.

Next the data were examined in terms of the CSC test score distribution and the effect of adding additional credit for veterans preference and aviation-related experience on the score distribution for those who passed. Figure 1 provides the distribution of raw (CSC battery) scores for the total applicant group. Figure 2 shows the score distribution for all ATC applicants who passed the CSC test battery, by raw score and after adding extra credit for veterans preference and OKT test scores. It is evident that the granting of extra credit for veterans preference and OKT significantly shifts the distribution to higher scores, with a total of 899 applicants scoring 90 or above (29% of those passing) compared to 500 (16% of those passing) before adding extra credit.

Figure 3 shows similar distributions for men and women separately. A significant shift to higher scores is apparent for the men, but there was no noticeable change for the women. Of the two extra credit factors, veterans preference accounted for more of the shift than OKT credit. For every woman who received extra credit for being a veteran, 16 men also

Table 12

Comparison of OKT scores and resulting experience credit
with percentages of applicants claiming Veterans Preference.
1978 applicants, N = 968.

OKT Score	Experience Credit	Veterans Preference				TOTAL
		No		Yes		
		N	Percent	N	Percent	
65-69	3 points	103	45	126	55	229
70-74	5 points	76	33	153	67	229
75-79	10 points	61	24	190	76	251
80+	15 points	26	10	233	90	259
		266	28	702	72	968

Table 13

Distribution of applicants who passed and failed the CSC Test
Battery by extra-point groups based on the OKT.
1978 applicants, N = 968.

Extra Points	Total	Failed CSC Battery		Passed CSC Battery Eligible for Extra Credit	
		N	Percent	N	Percent
15 points	259	58	22	201	78
10 points	251	74	29	177	71
5 points	229	62	27	167	73
3 points	229	78	34	151	66
	968	272	28	696	72

429.
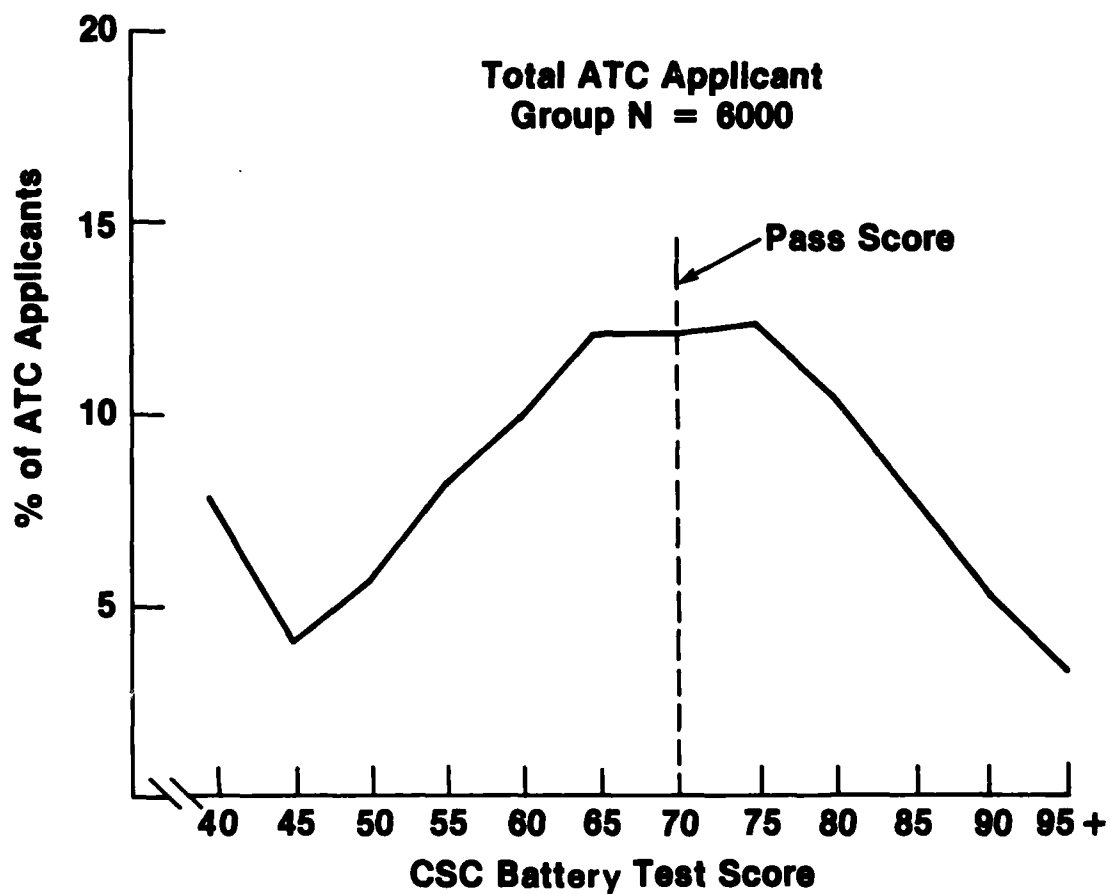
Figure 1.
Distribution of Raw Scores on the CSC Test Battery.
1978 Applicants, N = 6000
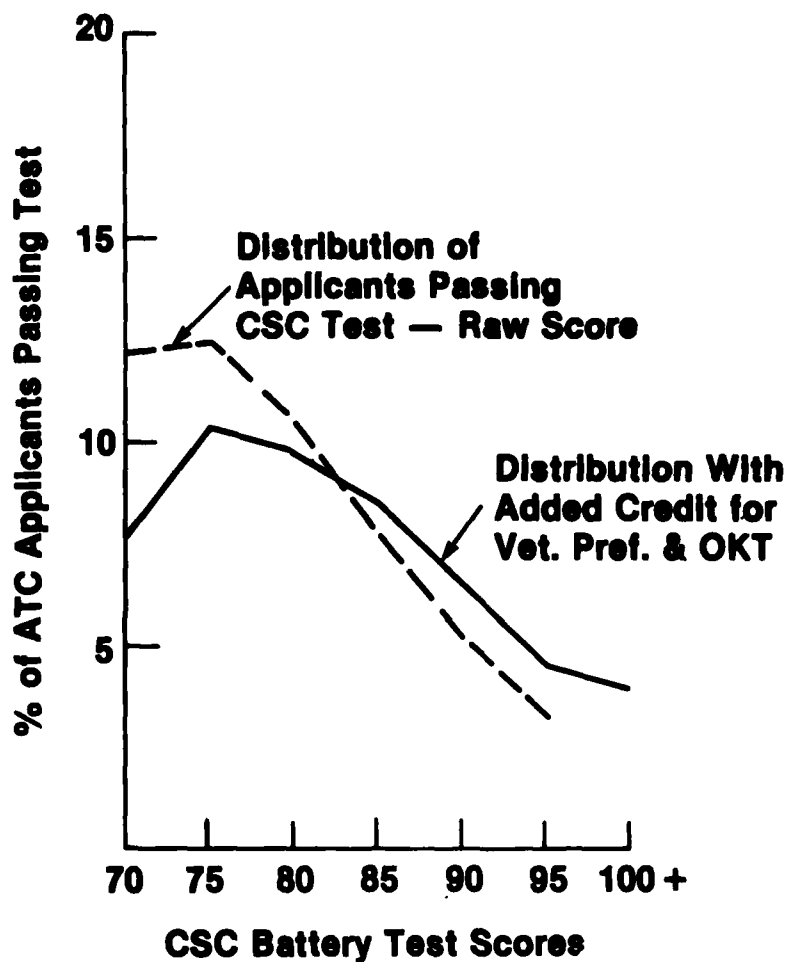
Figure 2.
**Distribution of CSC Battery Raw Scores (Above) and of
Scores With Extra Credit Added for Veterans
Preference and OKT Scores for All 1978 Applicants
Who Passed the CSC Test Battery.**

431.

Men
(N = 2236)

Women
(N = 799)

Distribution of
Men Passing Test
(Raw Score)

Distribution of
Men With Added
Credit for Vet.
Pref. & OKT

% of Applicants Passing

OPM ATC Test Scores

Distribution of
Women Passing Test
(Raw Score)

Distribution of
Women With Added
Credit for Vet. Pref.
& OKT

% of Applicants Passing
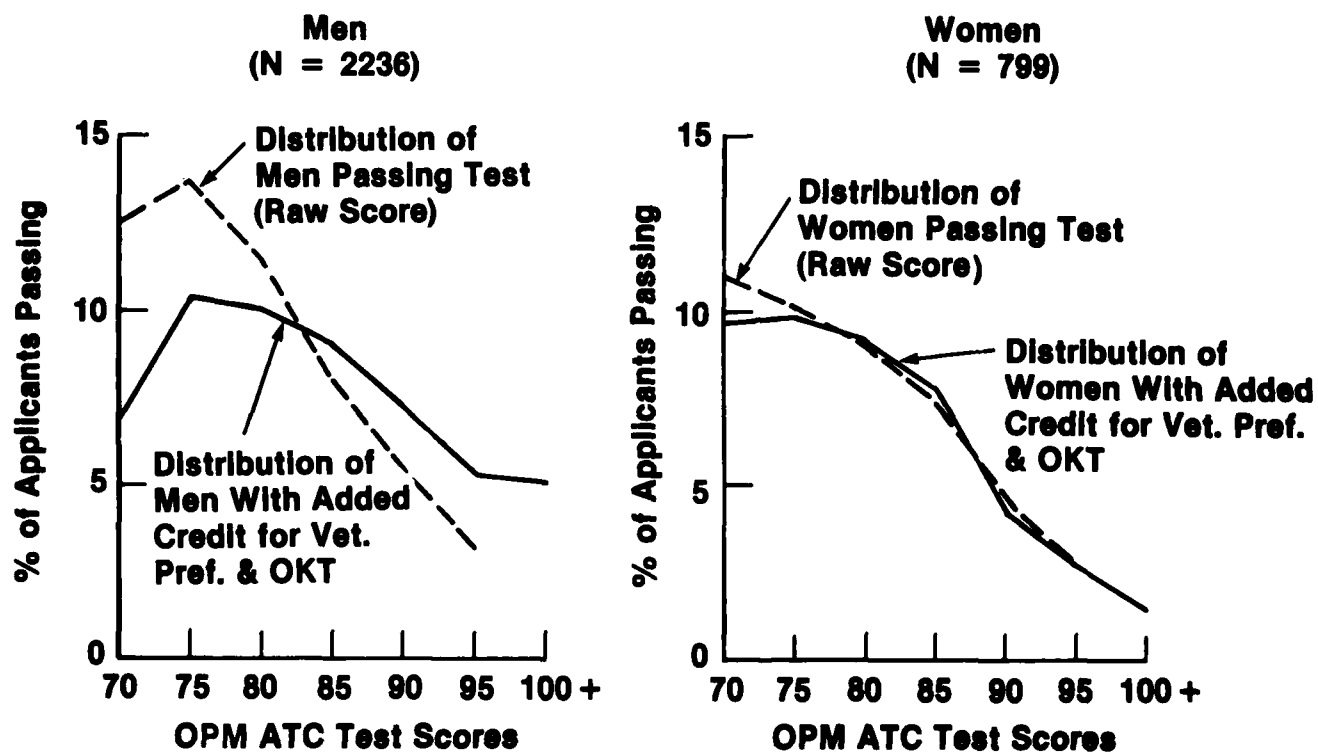
OPM ATC Test Scores

Figure 3.
Distribution of CSC Battery Raw Scores and of Scores
With Extra Credit Added for Veterans Preference and
OKT Test Scores, for Men and Women.
1978 Applicants Who Passed the CSC Test Battery.

432.

received veterans preference credit. By comparison, for every woman who would have received extra credit for aviation-related experience based on OKT scores, 9 men would also have received extra credit. While both factors resulted in moving more men than women to higher total score ranges for appointment consideration, almost twice as many men were moved up based on veterans preference than on OKT scores.

The effect on score distribution of granting extra credit can also be examined in terms of the relative representation of men, women, and minority groups in various score ranges, with and without the extra credit. Table 14 shows the percentage passing of men and women and of three race-ethnic groups in each of four score range groups, based on raw CSC test scores (without veterans preference or OKT extra credit) and compares those to the percentage of persons in each group after applying extra credit. It should be noted that by adding extra credit it was possible to have a maximum total score that exceeded 100.

Based on the raw CSC test scores, women comprised 26% to 27% of each of the score range groups. Since women comprised 26% of the total pass group, they were also proportionately represented (compared to men) in each of the score range groups. However, when veterans preference and OKT credit were included, the percentage of women increased to 33% in the 70-79 score group, remained the same in the 80-89 group, and dropped to 21% and 11% respectively, in the two highest scoring groups. These data provide a quantitative measure of the extent to which veterans preference and aviation-related experience, as measured by OKT, differentially affect men and women for appointment consideration.

In the case of the black and Hispanic minority groups, it is evident that the granting of veterans preference and OKT credit had very little impact on proportionate representation within each CSC score group for those applicants who passed the CSC test. Overall, these data indicate that veterans preference and credit for aviation experience affected each of the race-ethnic groups in essentially the same manner with respect to their score distributions.

Analysis of the Experimental Battery

Table 15 compares the numbers of applicants who passed and failed the CSC test battery and the experimental ATC battery; for this purpose, test records were available for 5976 of the total sample of 6000 applicants. To "pass" the experimental battery, an applicant had to have a minimum raw score of 100.47, which equated to a mean score at approximately the 50th percentile; the OKT was not included in this score.

Table 15 shows that 46% of the applicants passed both tests and failed both. Of the remaining 12%, 7% passed the experimental battery and failed the CSC battery, while 5% passed the CSC and failed the experimental battery. The 437 applicants who passed the experimental battery but would not have been eligible for appointment based on the CSC

433.

Table 14

Percentage passing the CSC test battery before and after adding
extra credit. Data are presented in four score ranges, for men
and women and for three race-ethnic groups. 1978 applicant sample.
N - pass sample, sex groups: 3035, pass sample, race-ethnic
groups included: 2948.

Group	70-79 Raw Score N=1440	70-79 With Credit N=1076	80-89 Raw Score N=1106	80-89 With Credit N=1086	90-100 Raw Score N=489	90-99 With Credit N=644	100+ With Credit N=229
Men	74%	67%	73%	73%	73%	79%	89%
Women	26%	33%	27%	27%	27%	21%	11%
	100%	100%	100%	100%	100%	100%	100%
White	81%	80%	90%	89%	94%	93%	93%
Hispanic	5%	5%	4%	4%	4%	4%	4%
Black	14%	15%	6%	7%	2%	3%	3%
	100%	100%	100%	100%	100%	100%	100%

Table 15

Percentages of 1978 applicants who passed and failed the CSC
and the experimental test batteries. N = 5976.

	Passed CSC	Failed CSC	TOTAL
Passed experimental battery	2727 (46%)	437 (7%)	3164 (53%)
Failed experimental battery	308 (5%)	2504 (42%)	2812 (47%)
TOTAL	3035 (51%)	2941 (49%)	5976 (100%)

435.

battery represent 14% of those who passed the experimental battery. Most of the applicants who failed both batteries fell around the passing score on both tests. However, there were some who scored as low as 45 on the CSC who passed the experimental battery, and others who scored as high as 85 on the CSC who failed the experimental battery. In an effort to understand these differences, pass-fail rates were computed by sex and race-ethnic groups for both tests. These are shown in Table 16; data for the CSC battery also appear in Table 7.

It appears that the pass rate on the experimental battery was 2% higher than on the CSC battery (53% vs 51%). This occurred because the mean raw score was used to identify "passes" on the experimental battery, instead of the established pass score, which was slightly higher. This difference is reflected in the pass-fail percentages for men and women. However, for the race-ethnic groups, the increases in percent of passes were slightly greater for whites and Hispanics, while the pass percent for blacks dropped from 19% to 18%. In view of the disporportionate number of minorities who failed the experimental battery, a breakdown of the score distribution was examined; this is shown in Table 17. It is evident from this analysis that, for all race-ethnic groups, the number of applicants who did very poorly on the Experimental battery (scored below 50) was much smaller than on the CSC battery (see Table 8). For example, 32% of the black applicants scored below 50 on the CSC battery, but only 17% on the experimental battery. This also held true for applicants who scored very high on the CSC battery, where 11% of the white group scored above 89, in contrast to only 2% on the experimental battery.

Table 18 compares the experimental battery with the CSC battery in relation to percentages of race-ethnic groups in the civilian labor force, the applicant sample, and the subgroups who passed each of the tests; this table incorporates data on the CSC battery from Table 9. Considering the use of mean scores to identify passes on the experimental battery, discussed above, no differences in proportions passed were discernible by race-ethnic group. However, the proportions of minorities who passed each of the tests compared favorably with their frequency in the labor force.

Eligibility for extra credit, based on the OKT. Since 12% of the applicants failed one of the two test batteries, as discussed earlier, the numbers eligible for extra credit on the experimental battery were different from those on the CSC battery. Table 19 shows the number and distribution by "point groups" of those in the total applicant group who were eligible for extra credit based on their OKT scores, who failed the experimental battery and those who passed. Of the total of 968 applicants who could have been eligible for OKT extra credit, a total of 206 failed both the CSC and the experimental battery. Sixty-six failed the CSC battery, but passed the experimental battery, and 68 passed the CSC battery, but failed the experimental battery. Consequently, the two test batteries affected 340 applicants, or 35 percent of the applicants who scored above 64 on the OKT.

436.

Table 16

Pass and fail percentages on the CSC and experimental batteries by sex and race-ethnic groups. 1978 applicants. N = 5976 (sex groups, 5813 (race-ethnic groups).

Group	No. of applicants	Pass				Fail			
		Exper. Battery		CSC Battery		Exper. Battery		CSC Battery	
		N	%	N	%	N	%	N	%
Men	4191	2312	55	2236	53	1879	45	1955	47
Women	1785	852	48	799	45	933	52	986	55
White	4067	2678	66	2556	63	1389	34	1511	37
Hispanic	339	141	42	128	38	198	58	211	62
Black	1407	256	18	264	19	1151	82	1143	81

437.

Table 17

Distribution of scores on the experimental battery for white, Hispanic, and black applicants. 1978 applicant sample. N = 5813.

Group	Pass Range						Fail Range						TOTAL
	89 & over		80-89		70-79		60-69		50-59		below 50		
White	63	2%	857	21%	1758	43%	1021	25%	329	8%	39	1%	4067
Hispanic	3	1%	33	10%	105	31%	101	30%	74	21%	23	7%	339
Black	1	-	28	2%	227	16%	441	32%	465	33%	245	17%	1407

Table 18

Percentages of race-ethnic groups in the civilian labor force, the 1978 applicant population (N = 6000), and the subgroups that passed the experimental test battery and the CSC battery.

Percentages in

Group	Civilian Labor Force	1978 Applicant Sample	Pass Group CSC Battery	Pass Group Exper. Battery
Am. Ind.	.3%	1%	1.0%	1.2%
Asian	.8%	1%	1.1%	1.3%
Black	9.7%	24%	8.9%	7.6%
White	85.0%	68%	84.8%	85.5%
Hispanic	4.2%	6%	4.2%	4.4%
	100.0%	100%	100.0%	100.0%

Table 19

Distribution of 1978 applicants who were eligible for extra credit based on OKT scores, who failed the experimental battery and of those who passed, by "point groups."

Extra Points	Total	Failed Exper. Battery		Passed Exper. Battery	
		N	Percent	N	Percent
15 pts.	259	60	23	199	77
10 pts.	251	70	28	181	72
5 pts.	229	59	22	170	78
3 pts.	229	85	37	144	63
	968	274	28	694	72

440.

Figure 4 shows the distribution of raw scores on the experimental and CSC test batteries. Applicants who scored below 45 on the CSC battery tended to cluster in the 50-59 score range on the experimental battery. Table 20 also shows the comparative distribution statistics for the pass group on both batteries. It is apparent that the experimental battery provided considerably greater differentiation, compared to the CSC battery, on test scores among the applicants wh passed. The experimental battery had a much greater clustering of applicants in the 70-79 score range (68% vs 47%), and a much smaller percentage of applicants in the 90 and over range (2% vs 16%).

Extra credit. With respect to the effects of adding extra credit for veterans preference and experience, as reflected by OKT scores above 64, the raw score distribution and the distribution after adding extra credit were derived, as shown in Figure 5. On the experimental battery, 485 applicants (15% of those who passed) scored 90 or higher after receiving veterans preference and OKT credit, compared to 69 (2%) based on raw scores. This is in contrast to the CSC battery, on which 899 (29%) scored 90 or above after receiving extra credit.

Figure 6 shows comparable distributions for men and women and Table 21, the percentages of men and women in each of four score ranges, based on experimental battery raw scores and scores after adding extra credit. The overall effect of adding extra credit based on veterans preference and OKT scores to the experimental battery raw scores was essentially the same as for the CSC battery (See Table 14); the proportion of women increased in the lower score range (70-79) and decreased substantially in the higher range (90 and over).

The addition of credit for veterans preference and OKT scores had a very slight impact on race-ethnic groups, even in the 70-79 score range, where the percentage of whites decreased by 3% (84% to 81%) and the percentage of blacks increased by 2% (11% to 13%). Data for the experimental battery are shown in Table 22, and comparable data for the CSC battery, in Table 14.

Effects of weighting the experimental battery. The analyses of the experimental test battery up to this point have been based on equal weighting of each of the three tests (CSC-24, CSC-157, and MCAT) included in the experimental test battery.

Multiple regression analysis, with ATC Academy laboratory problem scores as the criterion, for a sample of 1827 ATC trainees, had resulted in a correlation (R) of .5407 (Boone, 1979). Converting the beta weights to unit weights of:

        1 x CSC 24 Arithmetic Reasoning
        2 x CSC 157 Abstract Reasoning and Letter Sequence
        4 x Multiple Controller Aptitude

441.
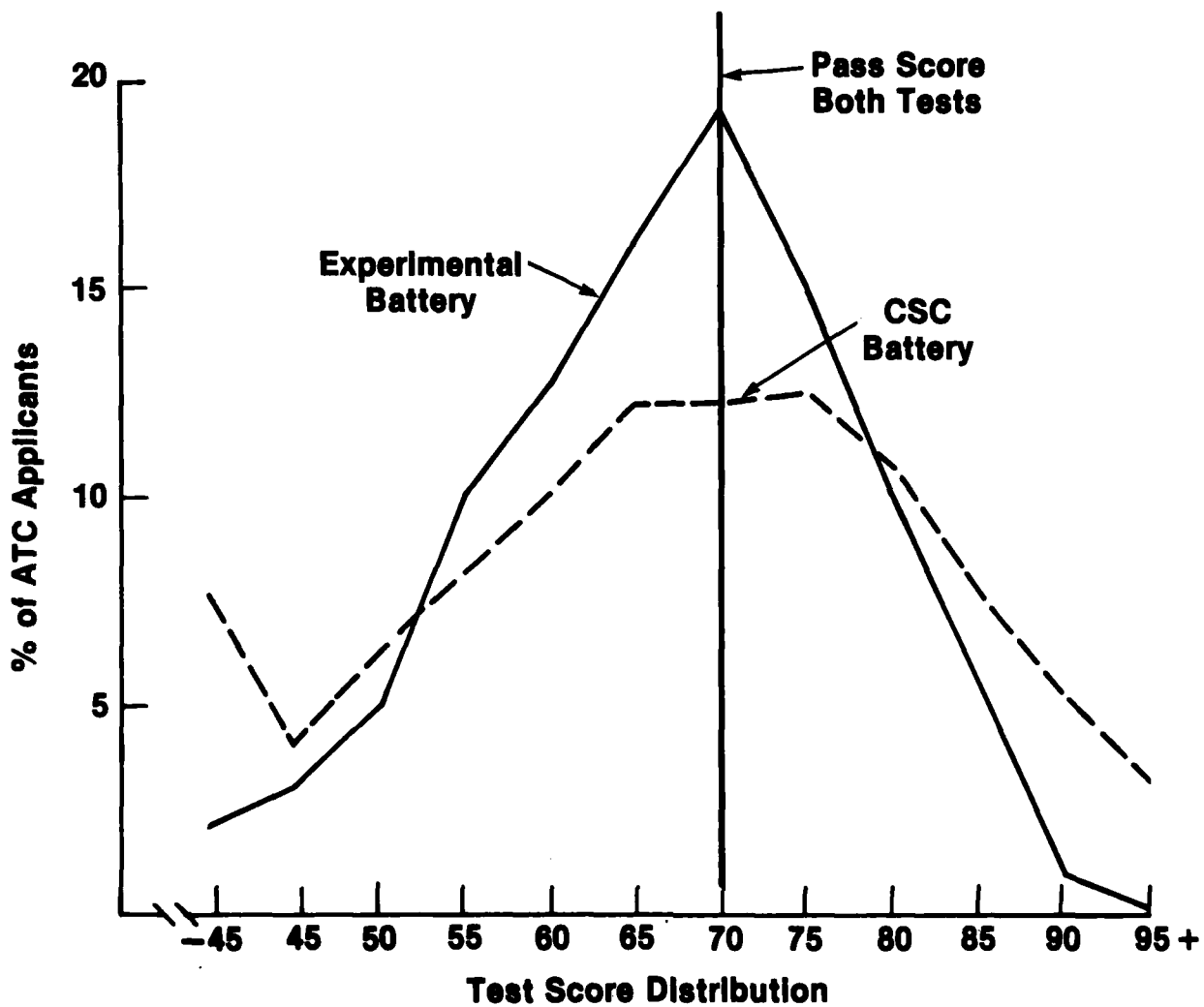
Figure 4.
Distribution of Raw Scores on the Experimental and CSC Test Batteries.
1978 Applicant Sample.

442.

Table 20

Pass group test score distribution on both batteries.
1978 applicant sample

	70-74		75-79		80-84		85-89		90-94		95 & over		TOTAL
	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	
CSC Battery	697	23	743	24	635	21	471	16	310	10	179	6	3035
Exper. Battery	1253	40	895	28	606	19	341	11	63	2	6	-	3164

443.
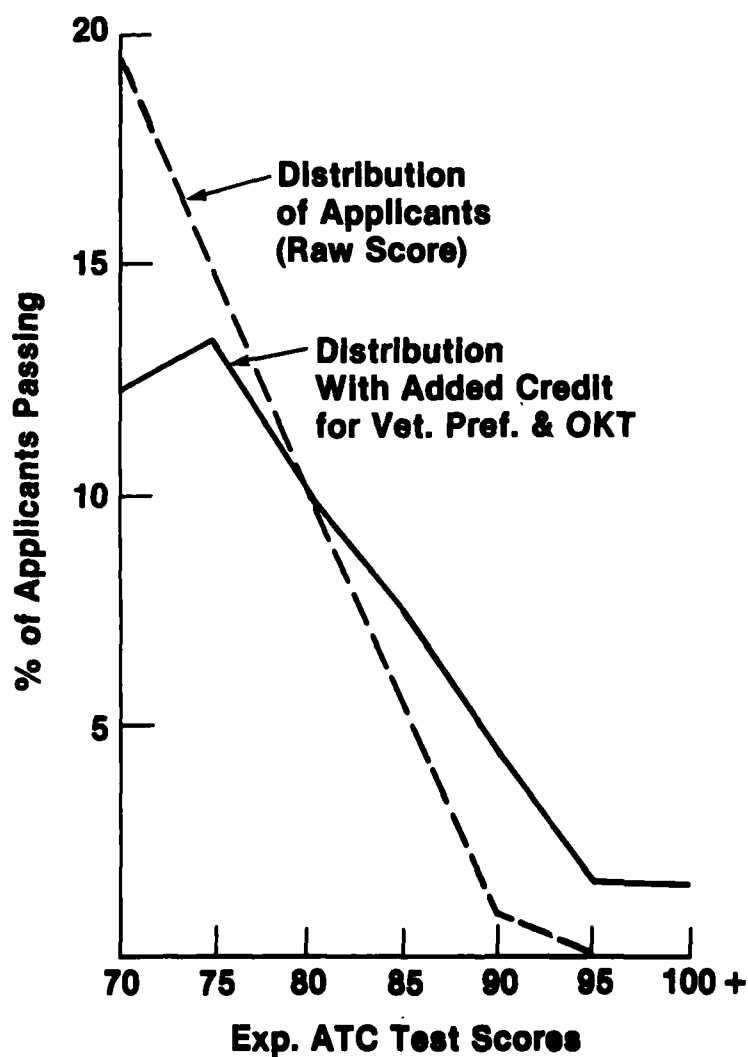
Figure 5.
Distribution of Raw Scores on the Experimental Battery for All Applicants Who Passed, and Scores After Adding Extra Credit for Veterans Preference and OKT Scores Above 64.

444.

Men
(N = 2312)

Women
(N = 852)

Distribution of
Men Passing Test
(Raw Score)

Distribution of
Men With Added
Credit for Vet.
Pref. & OKT

Distribution of
Women Passing Test
(Raw Score)

Distribution of
Women With Added
Credit for Vet. Pref.
& OKT

% of Applicants Passing

EXP ATC Test Scores
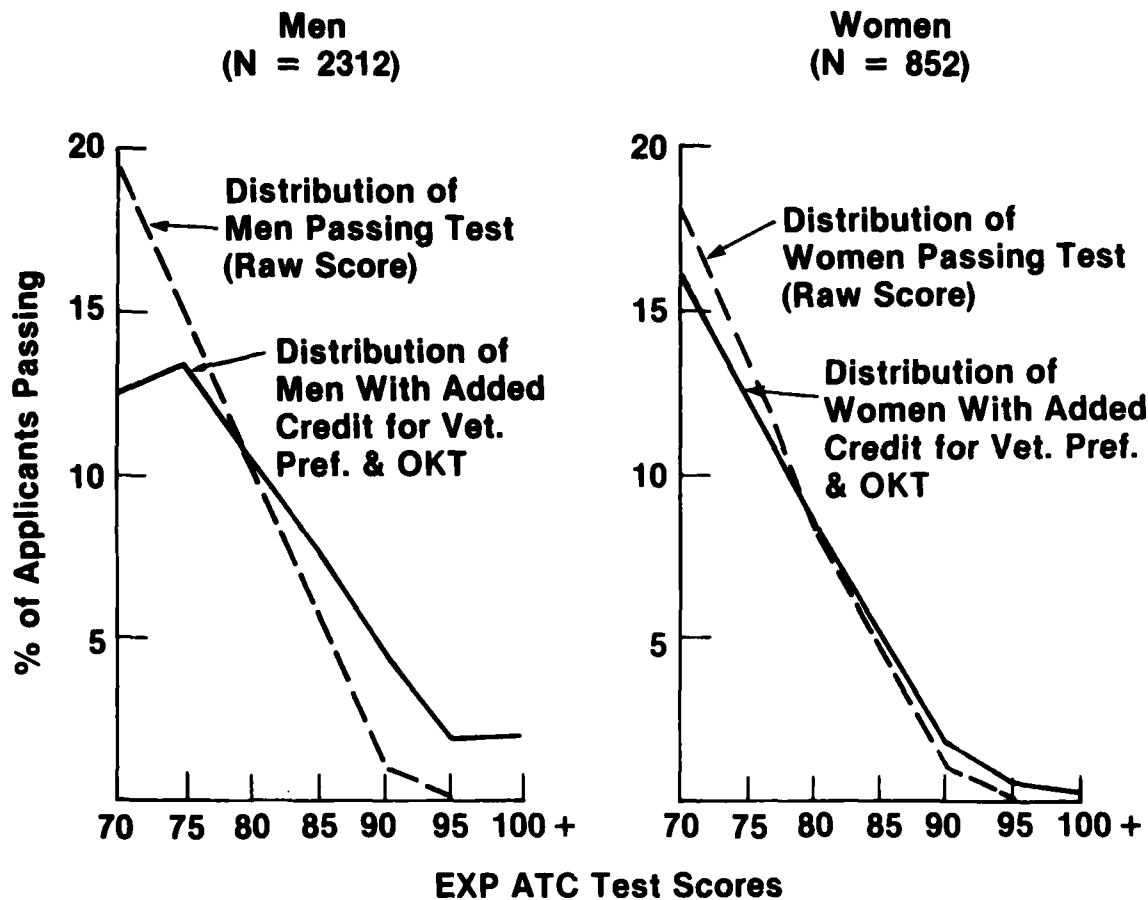
Figure 6.
Distribution of Raw Scores for Men and Women Applicants Who Passed the Experimental Battery and of Scores After Adding Extra Credit for Veterans Preference and OKT Scores Above 64.

445.

Table 21

Percentages of men and women applicants in each of 4 score ranges on the experimental battery, before and after adding extra credit for veterans preference and OKT scores over 64.

| | 70-79 | | 80-89 | | 90-100 | | 100+ |
	Raw Score N=2148	With Credit N=1523	Raw Score N=947	With Credit N=1156	Raw Score N=69	With Credit N=374	With Credit N=111
Men	72%	65%	75%	76%	71%	89%	95%
Women	28%	35%	25%	24%	29%	11%	5%
	100%	100%	100%	100%	100%	100%	100%

446.

Table 22

Percentages of white, Hispanic, and black applicants who passed the experimental battery, in 4 score ranges, before and after receiving extra credit for veterans preference and OKT scores. 1978 applicant sample (pass only, N = 3075).

	70-79		80-89		90-100	90-99	100+
	Raw Score N=2090	With Credit N=1475	Raw Score N=918	With Credit N=1127	Raw Score N=67	With Credit N=364	With Credit N=109
White	84%	81%	93%	92%	94%	94%	93%
Hispanic	5%	6%	4%	4%	5%	3%	5%
Black	11%	13%	3%	4%	1%	3%	2%
	100%	100%	100%	100%	100%	100%	100%

447.

resulted in a correlation (R) of .5354. In order to assess further the effect of using the weighted test scores, the 1978 ATC applicant group test results were analyzed, based on a weighted combination of the three scores.

As shown in Table 23, with the weighted composite score, the percentage of men who passed increased from 55% to 56%, while the percentage of women who passed decreased from 48% to 43%. The pass rate for whites remained the same (66%); Hispanics dropped from 42% to 41%, and blacks dropped from 18% to 14%.

Related shifts in the score distribution for race-ethnic groups, attributable to weighting, are shown in Table 24, which uses data from Table 17. Although the shifts were not great, there was a tendency for whites to shift toward higher scores, particularly in the 80-89 range, for Hispanics to remain unchanged, and for blacks to shift to lower scores, particularly in the fail range. For the sample as a whole, the number of scores below 50 shifted from 307 (5.3%) - unweighted, to 229 (3.5%) - weighted. The distribution of weighted experimental battery scores is shown graphically for the three race-ethnic groups, in Figure 7.

Comparison of the weighted experimental battery scores and the CSC battery scores in terms of pass-fail status of the applicants showed that: (1) 2606 passed both tests, (2) 2431 failed both bothes, (3) 510 passed the weighted experimental battery but failed the CSC battery, and (4) 429 passed the CSC battery but failed the weighted experimental battery. Use of these two different batteries affected the pass-fail status of 939 (16%) of the applicants. With the weighted experimental battery, 510 different applicants (16% of those who passed) would have been eligible for appointment consideration who would not have been eligible based on the CSC battery.

Comparison of the pass-fail results for the weighted and unweighted experimental battery scores showed that 8% of the total applicants were affected; 219 failed the unweighted battery but passed the weighted battery and 267 passed the unweighted but failed the weighted battery.

In relation to the effect of weighting on representation of race-ethnic groups among those who passed the experimental battery, Table 25, which uses data from Table 18, shows percentages of five groups in the civilian labor force, the 1978 applicant sample, and the subgroups that passed the CSC battery, the experimental battery (unweighted), and the experimental battery (weighted). The differences between the latter 2 were slight; the largest were the decrease of blacks from 7.6% to 6.4%, and the increase of whites from 85.0% to 87.1%.

Comparison of the weighted experimental battery scores and the CSC battery scores, both including extra credit for veterans preference and OKT scores, is shown in Figure 8.

448.

Table 23

Comparison of percentages of 1978 applicants who passed and failed
the experimental battery, by sex and race-ethnic groups.
N = 5976 (sex groups) and 5813 (race-ethnic groups).

Group	N	Pass Weighted N	Pass Weighted Percent	Pass Unweighted N	Pass Unweighted Percent	Fail Weighted N	Fail Weighted Percent	Fail Unweighted N	Fail Unweighted Percent
Men	4191	2348	56	2312	55	1843	44	1879	45
Women	1785	768	43	852	48	1017	57	933	52
TOTAL	5976	3116	52	3164	53	2860	48	2812	47
White	4067	2698	66	2678	66	1369	34	1389	34
Hispanic	339	136	41	141	42	203	59	198	58
Black	1407	198	14	256	18	1209	86	1151	82
TOTAL	5813	3032	52	3075	53	2781	48	2738	47

Table 24

Distribution of weighted (w) and unweighted (un) scores on the experimental battery for white, Hispanic, and black applicants. 1978 applicant sample.
N = 5813

Group		89 & over	Pass Range 80-89	70-79	60-69	Fail Range 50-59	Under 50	TOTAL
White	w	66 (2%)	987 (24%)	1645 (40%)	1076 (27%)	273 (7%)	20 (<1%)	4067
	un	63 (2%)	857 (21%)	1758 (43%)	1021 (25%)	329 (8%)	39 (1%)	
Hispanic	w	4 (1%)	32 (9%)	100 (30%)	100 (30%)	65 (19%)	22 (7%)	339
	un	3 (1%)	33 (10%)	105 (31%)	101 (30%)	74 (21%)	23 (7%)	
Black	w	0 (0%)	31 (2%)	167 (12%)	500 (36%)	522 (37%)	187 (13%)	1407
	un	1 (0%)	28 (2%)	227 (16%)	441 (32%)	465 (33%)	245 (17%)	

450.
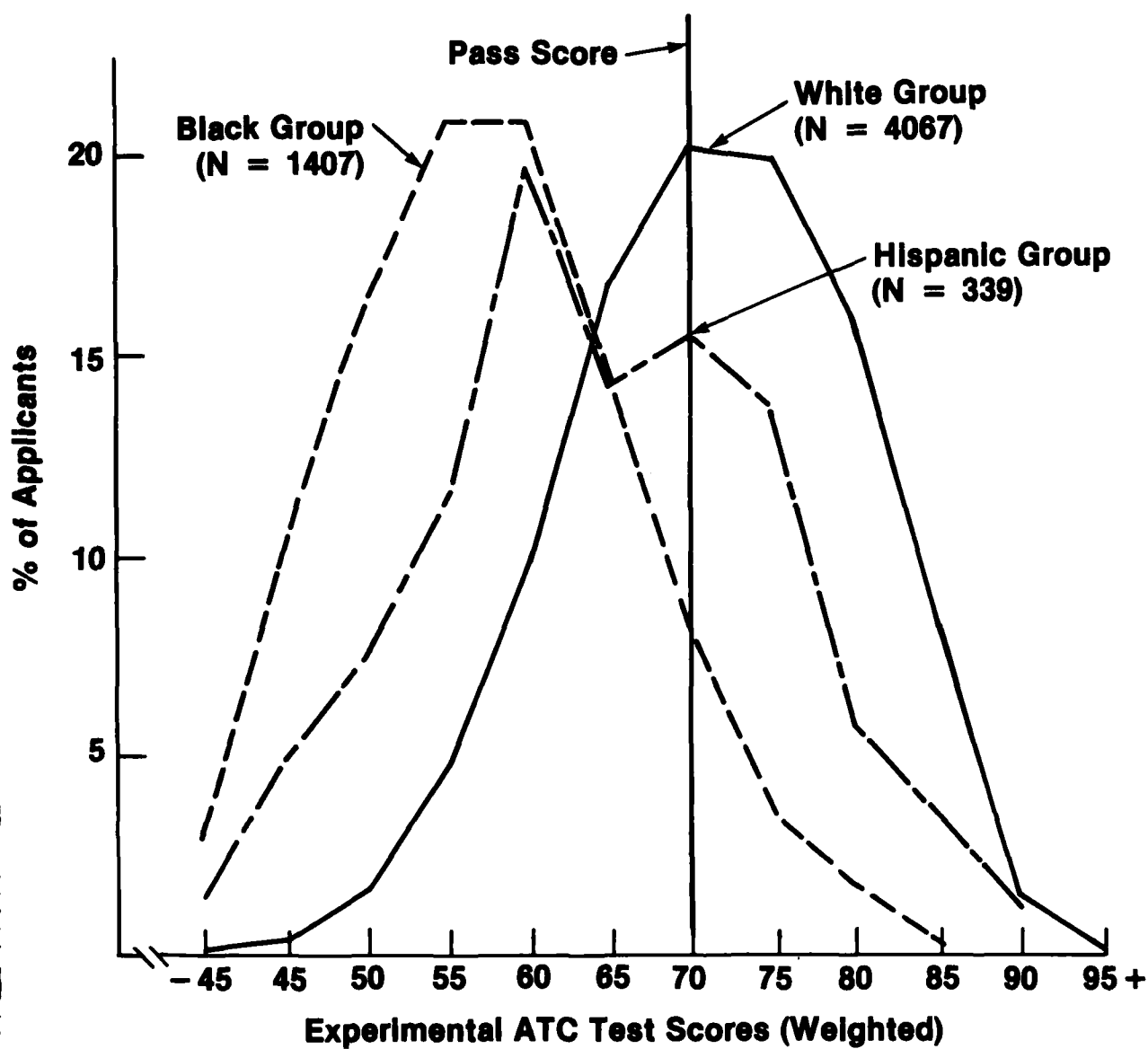
Figure 7.
Distribution of Weighted Experimental Battery Scores for White, Hispanic, and Black Applicants.

451.

Table 25

Percentages of race-ethnic groups in the civilian labor force,
the 1978 applicant population (N = 6000), and the subgroups
that passed the CSC battery and the experimental battery --
unweighted and weighted.

Percentages in

Group	Civilian Labor Force	1978 Applicant Sample	CSC Battery Pass Group	Experimental battery pass group	
				unweighted	weighted
American Indian	.3	1.0	1.0	1.2	1.1
Asian	.8	1.0	1.1	1.3	1.0
Black	9.7	24.0	8.9	7.6	6.4
White	85.0	68.0	84.8	85.5	87.1
Hispanic	4.2	6.0	4.2	4.4	4.4
	100.0	100.0	100.0	100.0	100.0

452.
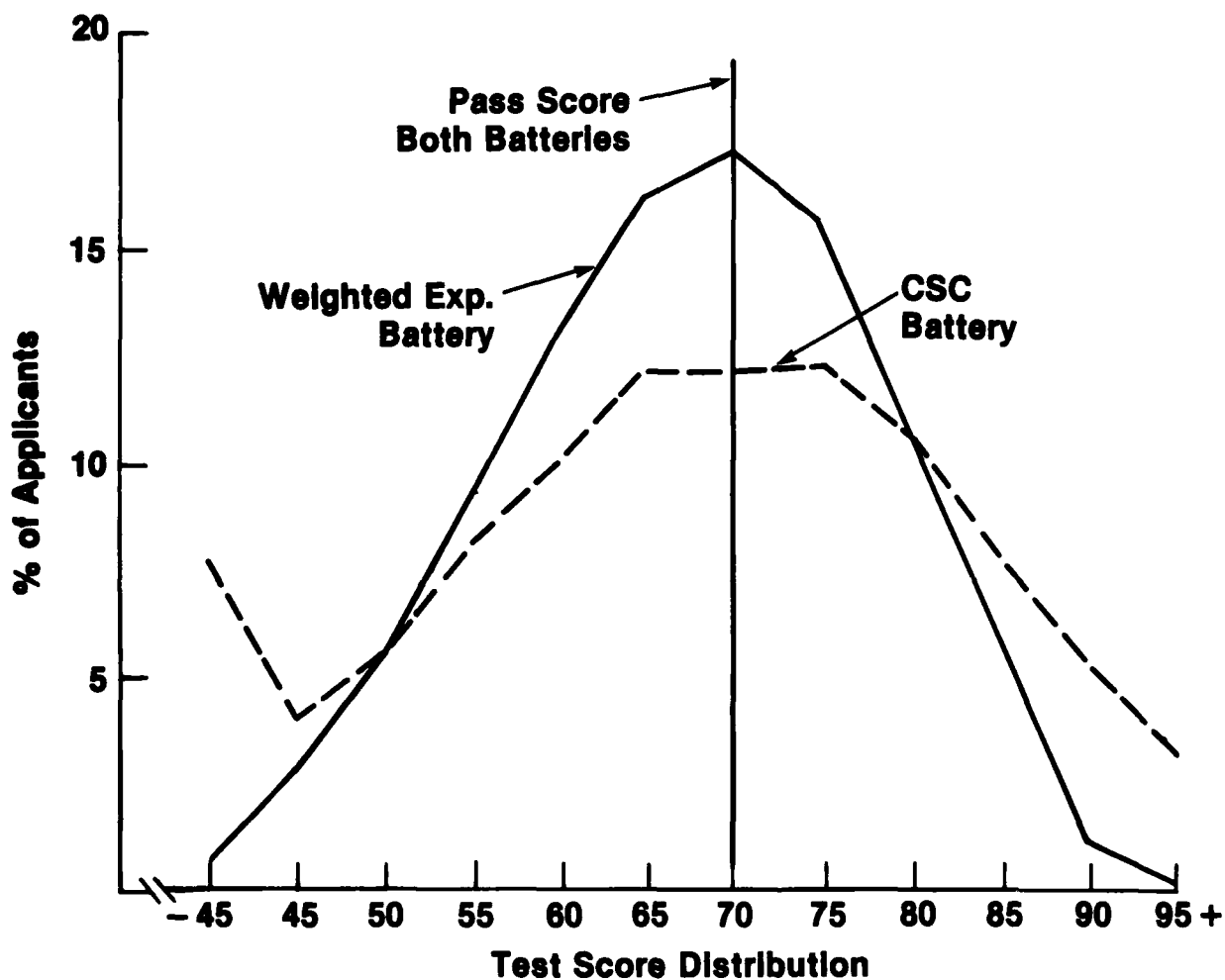
Figure 8.
Distribution of Scores (Including Extra Credit for Veterans Preference and OKT Scores) on the CSC and Experimental (Weighted) Batteries.

453.

The effect of adding extra credit for veterans preference and OKT scores to the weighted experimental battery scores, compared to the unweighted scores, is shown in Table 26 for sex and race-ethnic groups. As implied in earlier discussion, when weights were used, men tended to shift to higher scores and women to lower scores; only two shifts larger than 1% were noted among the race-ethnic groups: in the 90-99 range, Hispanics decreased from 6% to 4% and blacks increased from 0% to 3%.

Tables 27 and 28 present summary data comparing the CSC battery and weighted and unweighted scores on the experimental battery. Table 27 shows score distributions in the pass range, with extra credit included, and pass rates for the three battery scales. Table 28 gives descriptive statistics on the three battery scales for sex and race-ethnic groups.

Finally, the intercorrelations of the three battery scores are shown in Table 29.

CONCLUSIONS

Given the high attrition rate of ATC trainees and the objective of developing improved methods of selecting applicants for the ATC occupation, attention was focused on the selection tests used in the hiring process. The results obtained with the experimental test battery, which was developed in a series of research studies discussed in Chapter 18, were not unexpected.

The number of applicants who scored high on the experimental battery (85 and above) was much smaller than on the CSC battery. Based on the validity studies reported in the following chapter, applicants who score high on the experimental battery can be expected to have very high probability of success in the developmental training for the ATC occupation.

Granting extra credit for military service is required by law. This does impact on the competitive appointment consideration for women, but does not appear to have a differential effect on individual minority groups. Aviation-related experience and knowledge, measured by the Occupational Knowledge Test, has been shown to be a positive predictor of success in ATC training (Chapter 16). Granting extra credit for the demonstration of this knowledge improves the competitive position of selected applicants, but this is consistent with the objective of developing improved selection procedures in order to hire those applicants who demonstrate aptitude for air traffic control work and whose potential for success is greatest.

Table 26

Comparison of weighted scores on the experimental battery, before and after adding extra credit for veterans preference and OKT scores, in 4 ranges of passing score, by sex and race-ethnic groups.

Percentages in

| Group | 70-79 | | 80-89 | | 90-100 | 90-99 | 100+ |
	Raw Score N=1968	With Credit N=1426	Raw Score N=1076	With Credit N=1162	Raw Score N=72	With Credit N=402	With Credit N=126
Men	74.	67.	77.	78.	88.	92.	96.
Women	26.	33.	23.	22.	12.	8.	4.
	N=1912	N=1381	N=1050	N=1136	N=70	N=392	N=123
White	86.	85.	94.	93.	94.	93.	94.
Hispanic	5.	6.	3.	3.	6.	4.	5.
Black	9.	9.	3.	4.	0.	3.	1.

455.

Table 27

Score distributions in the pass range (extra credit included) for the CSC battery, the unweighted experimental battery, and the weighted experimental battery. 1978 applicants, N = 5976.

Test Battery	70-79		80-89		90-99		100+		Total Pass		All Applicants
Earned Rating	N	Pct.	N	Pct.	N	Pct.	N	Pct.	N	Pct.	N
CSC Battery	1076	35	1086	36	644	21	229	8	3035	51	5976
Exper. Battery - un	1523	48	1156	36	374	12	111	4	3164	53	5976
Exper. Battery - w	1426	46	1162	37	402	13	126	4	3116	52	5976

456.

Table 28

Means and Standard Deviations for CSC Battery and Experimental
Battery (Unweighted and Weighted) for subgroups of 1978 applicants
by sex and race-ethnic groups.

Group	CSC Battery		Exp. Battery - Un		Exp. Battery - W	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Men (N=4191)	210.0	46.0	101.9	22.6	227.3	55.5
Women (N=1785)	199.8	49.3	98.3	22.7	210.7	56.1
White (N=4067)	220.1	39.8	107.5	19.0	241.5	46.7
Hispanic (N=339)	191.7	49.7	93.9	23.4	206.5	56.1
Black (N=1407)	165.0	45.8	79.6	21.0	170.5	47.2
Asian (N=57)	210.7	48.3	103.8	21.7	235.8	50.8
Amer. Indian (N=61)	204.5	41.6	99.2	20.5	222.9	49.6
Total Group	207.1	47.2	100.5	22.7	222.3	56.2
Pass Score	210.0	-	100.47	-	222.27	-
Transmuted Score	70.0		70.0[1]	-	70.0[2]	-

[1] Unweighted experimental battery

$$Ts = 10.56(\frac{Rs-Rs\overline{X}}{RsSD}) + 70 \quad \text{where:}$$

Rs = Raw score on unweighted ATC test
$Rs\overline{X}$ = Mean of unweighted ATC test (100.47)
RsSD = Standard deviation of unweighted test battery (22.67)

[2] Weighted experimental battery

$$Ts = 10.68(\frac{Rs-Rs\overline{X}}{RsSD}) + 70 \quad \text{where:}$$

$Rs\overline{X}$ = 222.27
RsSD = 56.22

Table 29

Intercorrelations of CSC battery and experimental battery -
unweighted and weighted scores (with extra credit included).

	CSC Battery	Exper. Battery - Un.	Exper. Battery - W.
CSC Battery	1.00	.92	.88
Exper. Battery - Un.		1.00	.97
Exper. Battery - W.			1.00

Chapter 21

VALIDITY AND UTILITY OF THE ATC EXPERIMENTAL TEST BATTERY

STUDY OF ACADEMY TRAINEES, 1978[1]

Donald B. Rock, John T. Dailey, Herbert Ozur,
James O. Boone, and Evan W. Pickrel

INTRODUCTION

Following the extensive developmental analyses described in the
preceding chapters, the next step was to validate the experimental
test battery on an independent sample of subjects. This was accom-
plished, using a new sample of ATC trainees attending the FAA Initial
Qualification Terminal and EnRoute training program, who were enrolled
between June and September, 1978. This study was summarized in the
report by Rock, Dailey, Ozur, Boone, and Pickrel (1982) and the major
findings are included in this chapter.

High validity was obtained, consistent with earlier research. In
addition, the superiority of the new (experimental) battery to the CSC
battery that it replaced in 1981 is demonstrated by the following tabu-
lation of fail rates in Academy training for the two test batteries:

Transmuted Score Range	Percentage Failure Rate in Academy Training	
	CSC Battery	Exper. Battery
95+	22.8	0.
90-94	31.8	18.5
85-89	39.7	20.4
80-84	45.0	28.0
75-79	50.0	42.2
70-74	56.4	51.0
Below 70	-	79.2

All of the trainees in this sample had previously passed the CSC battery
to qualify for enrollment in the Academy. Nevertheless, there were 48
(9.2%) of these who scored below 70 (the passing grade) on the experimen-
tal battery and would not have been admitted. Of these, 38 (79.2) failed
the program. At all other score levels, in the passing range, the failure
rate on the experimental battery was lower than on the CSC battery, and
substantiall so in the range above 75.

DESCRIPTION OF THE STUDY

Study Sample

The sample of trainees included in this validation study consisted
of 953 new trainees attending the FAA Academy during the period June 1978

[1]Prepared by S. B. Sells

through December 1978. As in previous studies, tests were administered
on a voluntary basis on the first day of attendance at the Academy. The
composition of the sample with respect to sex and racial groups is shown
in Table 1, and the present sample is compared with the 1978 applicant
sample (discussed in the preceding chapter), in Table 2, which shows the
subgroups that passed the CSC battery (with a passing score of 70) and
that passed the CSC battery with a total score (including extra credit)
of 90 or over.

As shown in Table 2, the representation of minorities in the trainee
sample was somewhat greater than would have been expected if all trainees
had been selected through competitive appointment processes from the OPM
registers. Because of the large number of applicants in relation to the
relatively small number of vacancies, competitive selection of ATC trainees
has normally been made in score ranges above 90. As shown in Table 2,
white applicants comprised 91% of all applicants who scored 90 or higher
on the CSC battery (on earned ratings including extra credit for veterans
preference and OKT scores). However, in the present sample of 953 trainees,
whites represented only 88% of the total. This discrepancy is accounted
for by the Predevelopmental Program discussed below.

Because of the difficulties in hiring women and minorities through
competitive appointment processes from OPM registers, FAA established an
alternate recruitment program in 1968 which provides for non-competitive
appointment of individuals who are already federal employees with status
in the career Civil Service system. These individuals are required to
pass the same ATC test battery with a score of 70 or more as the appli-
cants appointed from the OPM register. However, having passed the test,
they can be hired at the GS-5 level non-competitively. This alternative
recruitment program is identified as the ATC Predevelopmental Program.

Approximately 200 ATC trainees have been hired each year through this
alternate recruitment program. This represents about 10% of annual new
ATC hires. These employees have received a year of predevelopment train-
ing in air traffic control related subjects, including 17 weeks of formal
classroom instruction at the Predevelopmental Training Center at the Uni-
versity of Oklahoma, followed by training assignments at FAA Terminal
and EnRoute centers, and Flight Service Station field facilities. Upon
completion of predevelopmental training, these employees have been pro-
moted to GS-7 (the normal entry grade for ATC trainees) and assigned to
the FAA Academy for Initial ATC Qualification Training.

In order further to understand the composition of the 1978 ATC trainee
group, Table 3 identifies the predevelopmental employees included in the
sample, by sex and race-ethnic group. It is evident that the Predevelop-
mental recruitment program has provided a major avenue for entry of women
and minorities into the GS-7 Initial Qualification ATC training program.
About 12% of this 1978 ATC trainee sample entered the ATC occupation
through the Predevelopmental Program. Approximately 35% of the women, 52%
of the Hispanics, and 68% of the blacks in this group of over 900 ATC
trainees came into the occupation through Pre-developmental recruitment
efforts.

Table 1

Sex and race-ethnic composition of the study sample of 953 Academy trainees -- June - December, 1978.

Group	Number	Percent
Men	800	85.0
Women	141	15.0
Not identified	12	-
White	839	88.0
Black	81	8.5
Hispanic	23	2.4
Asian	10	1.1

461.

Table 2

Comparison of 1978 trainee sample with the 1978 applicant sample, for sex and race-ethnic groups, by subgroups of the total applicant sample, those who passed the CSC battery, and those who earned 90 or higher on the CSC battery.

Group	1978 Applicant Sample						1978 Trainee Sample	
	Total Applicants		Passed CSC Battery		Passed CSC Battery, Score 90+			
	N	Pct.	N	Pct.	N	Pct.	N	Pct.
Men	4191	70	2236	74	722	82.6	800	85
Women	1785	30	799	26	151	17.2	141	15
White	4067	68.6	2556	84.8	791	91.0	839	88.0
Black	1407	23.7	264	8.8	26	3.0	81	8.5
Hispanic	339	5.7	128	4.2	31	3.6	23	2.4
Asian	57	1.0	35	1.2	15	1.7	10	1.1
Amer. Indian	61	1.0	31	1.1	6	.7	0	0

Table 3

Composition of the 1978 trainee sample by non-Predevelopmental and Predevelopmental trainees and by sex and race-ethnic groups.

Group	Non-Predevelopmental, Competitive Appointees			Predevelopmental, Non-Competitive Appointees			Total	
	N	Pct. of Group	Pct. of Total	N	Pct. of Group	Pct. of Total	N	Pct.
Men	739	88.9	92.4	61	55.5	7.6	800,	100.
Women	92	11.1	65.2	49	44.5	34.8	141	100.
	831	100.0	88.3	110	100.0	11.7	941	100.
White	797	94.5	95.0	42	38.1	5.0	839	100.
Black	26	3.1	32.1	55	50.0	67.9	81	100.
Hispanic	11	1.3	47.8	12	10.9	52.2	23	100.
Asian	9	1.1	90.0	1	1.0	10.0	10	100.
	843	100.0	88.5	110	100.0	11.5	953	100.

463.

Table 4 compares the 1978 trainee sample with the Civilian Labor Force (CLF) in relation to trainees hired competitively (non-predevelopmental) and noncompetitively (predevelopmental). The total sample differs from the CLF in that it is higher in the percentage of men and lower in the percentage of women. In the non-predevelopmental subsample, the black and Hispanic minorities and the women show a large excess over their strength in the CLF, as intended.

Predictor Tests

Since the study sample had already taken the CSC battery, the scores on the total battery and the two tests from that battery that were retained in the experimental battery (CSC-24 - Arithmetic Reasoning and CSC-157 - Abstract Reasoning) were to be retrieved from FAA records. The additional tests administered at the FAA Academy were:

> MCAT (4o6e, 4e6o, 6o7e, 6e7o, 7o4e, and 7e4o)
> OKT 101B (100 times), 101C (60 items, and
> 102 A, B, C, D, E, F, G, H (80 items keyed in each).

In this study, two of the parallel forms of MCAT were administered to each student. MCAT 1 was the designation of the first form administered in each case, and MCAT 2, of the second form. The intercorrelations of raw scores for each half- or part-test on each of two parallel forms of MCAT administered to 617 ATC students at the Academy are shown in Table 5.

The correlation between variables 3 (MCAT 1 Total) and 6 (MCAT 2 Total) in Table 5, which is .60, represents the test-retest correlation between comparable 2-part, 35 minute forms of MCAT, and is equivalent to a reliability of .75 for the combined score for this restricted group. In an unrestricted population (e.g. applicants) the same reliability coefficient would be .88. About 265 trainees in the total sample (of 953) were administered form 101B or 101C of the OKT; for purposes of analysis in the present study, these test scores were converted to a scale common to the eight versions of Form 102 of the OKT.

Means and standard deviations of scores on the tests administered and on the two CSC tests for which scores were retrieved are shown in Table 6; note that the N's for the CSC tests were only 592, indicating missing data on these for 38% of the sample. The parameters observed in this table are closely similar to those reported for the same tests in earlier studies.

Since about 12% of the 953 ATC trainees entered the training program through the Predevelopmental Program, it was of interest to establish the extent to which mean scores and standard deviations on each of the tests differed between the predevelopmental and non-predevelopmental groups. Table 7 provides the descriptive statistics for each of these groups by sex. It is important to recognize that the scores for the predevelopmental trainees on MCAT and OKT were obtained after they had completed a year of training. Scores on the CSC-24 and CSC-157 tests were obtained for both groups based on the CSC test battery administered prior to employment in the ATC occupation.

464.

Table 4

Percentage composition of sex and race-ethnic groups in the 1978 trainee sample in comparison to the Civilian Labor Force, by Non-Predevelopmental and Developmental components.

Group	Civilian Labor Force	Non-Predevelopmental Trainees	Predevelopmental Trainees	Total Sample
Men	62.0	89.0	55.0	85.0
Women	38.0	11.0	45.0	15.0
White	85.0	94.5	38.0	88.0
Black	9.7	3.1	50.0	8.5
Hispanic	4.2	1.3	11.0	2.4
Asian	.8	1.1	1.0	1.1
Amer. Indian	.3	0	0	0

Table 5

Intercorrelations of MCAT raw scores (number right) for each half- or part-test on each of two parallel forms administered to 617 1978 trainees.

	1	2	3	4	5	6
First test (MCAT 1)						
1. First half	-	.58	.89	.48	.42	.51
2. Second half		-	.89	.48	.49	.56
3. Total			-	.89	.89	.60
Second test (MCAT 2)						
4. First half				-	.52	.85
5. Second half					-	.89
6. Total						-

466.

Table 6

Means and standard deviations of men, women, and race-ethnic groups on the tests of the experimental battery and the OKT, 1978 trainee sample.

Test	MEN			WOMEN			TOTAL		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
MCAT 1	800	37.4	6.9	141	34.1	7.8	953	36.9	7.1
MCAT 2	800	42.9	5.8	141	40.7	6.4	953	42.6	6.0
CSC 24	515	46.6	6.5	67	47.2	7.1	592	46.6	6.6
CSC 157	515	38.8	6.2	67	39.8	6.0	592	38.8	6.2
OKT	800	64.9	15.5	141	57.7	15.1	953	63.9	15.6

Test	ASIAN			HISPANIC			BLACK			WHITE		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
MCAT 1	10	36.2	4.7	23	37.1	6.6	81	29.8	7.9	839	37.6	6.7
MCAT 2	10	43.2	5.7	23	42.7	7.4	81	36.6	6.8	839	43.2	5.5
CSC 24	1	44.0	-	7	48.7	6.3	39	44.2	7.1	545	46.9	6.5
CSC 157	1	43.0	-	7	38.7	5.4	39	35.3	6.1	545	39.1	6.2
OKT	10	61.3	13.5	23	59.0	19.9	81	66.2	13.1	839	63.9	15.8

467.

Table 7

Means and standard deviations of men and women in the
Non-Predevelopmental and Predevelopmental Groups of
1978 trainees on tests in the experimental battery
and OKT.

| | Non-Predevelopmental ATC Trainees | | | | | | Predevelopmental ATC Trainees | | | | | |
| | Men | | | Women | | | Men | | | Women | | |
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
MCAT 1	739	37.8	6.6	92	35.2	7.4	61	32.3	8.1	49	32.4	7.9
MCAT 2	739	42.3	5.4	92	41.5	6.2	61	38.3	8.1	49	39.5	6.8
CSC-24	495	46.7	6.5	48	48.1	7.1	20	44.3	6.6	19	45.0	6.7
CSC-157	495	38.9	6.2	48	40.8	5.7	20	35.6	5.7	19	37.3	6.2
OKT	739	64.4	15.7	92	52.0	15.6	61	72.5	10.8	49	66.1	9.4

468.

On all tests except OKT, the mean score of the predevelopmental group was lower than that of the non-predevelopmental group. On OKT it is evident that the year of predevelopmental training for trainees who were hired non-competitively was successful in providing knowledge on air traffic rules, regulations, and procedures, which the OKT was designed to measure. Both men and women predevelopmentals scored significantly higher than their non-predevelopmental counterparts as reflected by the mean scores on OKT for the two groups. In addition, the standard deviations on OKT for the predevelopmental group were considerably lower, indicating a much more restricted range of test scores.

On the two CSC tests (157 and 24), women in both groups scored somewhat higher than men. This is consistent with the pattern derived from the 1976-1977 ATC applicant group (Chapter 19) and the 1978 ATC applicant group, discussed in the preceding chapter, who passed the CSC battery.

Men and women in both groups increased their mean test scores on the second administration of the MCAT. However, the difference in mean scores between the predevelopmental and the non-predevelopmental groups remained fairly constant for each of the tests. Men in the predevelopmental program scored between four and five points lower and predevelopmental women between two and three points lower on each of the tests, compared to non-predevelopmental trainees.

Table 8 presents means and standard deviations for the pre-developmental and non-predevelopmental trainees by race-ethnic group. As a group, the predevelopmental trainees scored lower on all tests in comparison with the non-predevelopmental trainees, except on OKT. The greatest differences in mean scores were associated with the MCAT, which also had greater score ranges for the predevelopmental group. These differences were generally evident within each of the race-ethnic groups, as well. On the OKT, predevelopmentals in each race-ethnic group generally scored significantly higher, with a more restricted range of scores, than their non-predevelopmental counterparts. However, it does not appear that the academic ATC knowledge acquired during the one year of training by the predevelopmental group carried over, in application to the MCAT. On the MCAT, there was a consistent learning pattern between the first and second administration of the test for all trainees, although the non-predevelopmental trainees generally maintained a constant difference in mean scores on both tests, compared to the predevelopmental trainees. This pattern supports the use of a double length MCAT as a means of measuring this "learning ability" among applicants.

Criterion-Performance Measure

The criterion measure adopted for test validation in the present study was pass vs fail (including withdrew) in the ATC Academy training course. Of the trainees who withdrew, those who did so for medical reasons or because of personal or family emergencies, were excluded from the sample,

469.

Table 8

Means and standard deviations of race-ethnic groups of 1978 trainees for Non-Predevelopmental and Developmental trainees.

Non-Predevelopmental ATC Trainees

	White			Black			Hispanic			Asian			Total		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
MCAT 1	778	37.7	6.7	24	30.8	6.4	11	41.7	4.7	8	36.4	5.2	821	37.5	6.7
MCAT 2	778	43.2	5.4	24	38.5	5.6	11	43.2	5.3	8	43.4	5.7	821	43.1	5.5
CSC 24	532	46.9	6.5	15	42.4	7.1	4	51.8	6.2	1	44.0	-	552	46.9	6.5
CSC 157	532	39.1	6.2	15	35.4	6.3	4	42.0	3.6	1	43.0	-	552	39.0	6.2
OKT	778	63.5	16.0	24	58.3	15.1	11	48.1	20.5	8	60.9	15.4	821	63.1	16.1

Predevelopmental ATC Trainees

	White			Black			Hispanic			Asian			Total		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
MCAT 1	43	36.4	6.9	56	29.5	8.3	12	32.9	5.2	1	34.0	-	112	32.6	8.1
MCAT 2	43	42.1	6.0	56	35.9	7.1	12	42.3	9.2	1	37.0	-	112	39.0	7.5
CSC 24	13	43.9	6.8	24	45.4	7.0	3	44.7	4.5	-	-	-	40	44.8	6.6
CSC 157	13	38.2	6.5	24	35.3	6.1	3	34.3	4.5	-	-	-	40	36.1	6.1
OKT	43	70.1	10.3	56	69.9	10.4	12	69.1	13.6	1	60.0	-	112	69.7	10.7

470.

since their training was interrupted through no cause of the student. In those cases, the student was allowed to reenter training after the medical situation or emergency was resolved.

Essentially all of the students who fail or withdraw do so in the laboratory phase of training which requires the trainees to demonstrate and apply their knowledge and skill to air traffic control problems conducted in a laboratory environment. The structure of the grading of the laboratory phase of training is provided in Figure 1. Estimated reliability for the six EnRoute control problems is .79, and .75 for the Terminal control problems, for students attending the FAA Academy during the period June-December 1978.

The ATC laboratory training phase consists of six operational ATC problems designed to replicate the non-radar air traffic control environment. Trainees conduct identical control problems within a common air traffic control sector, using identical equipment and evaluation criteria for each ATC option. These operational problems provide stimulus and decision-making conditions representative of the job environment and require responses from the trainees in terms of actions that are both possible and required to control air traffic on-the-job with a minimum of restrictions. Students are given a series of practice problems before each of the six graded problems, as part of the training and learning process. Each graded operational problem takes about one hour and is scored by an instructor using established grading standards for defined types of student errors or deficiencies. Additionally, each student s performance is evaluated by the instructor on a rating scale of 40 to 100. Each problem is scored and evaluated by a different instructor.

After the operational problems, a Controller Skills Tests (CST, See Chapter 12) is administered. This test was designed to measure the application of knowledge and skills taught during the first months of training. Three basic elements for evaluation are distributed within the test: (1) application of aircraft separation standards -- students must respond to control situations presented by flight strips and charts; (2) responding to or forwarding information received, which pertains to coordination of aircraft movement or information with other controllers; and (3) other ATC control items, such as board management, timeliness of actions, and phraseology. The CST is a 100-item multiple choice test with an administration time of one hour. The correlation between the average score on the six operational problems and the CST for this group of trainees was .50 for those in the Terminal option (N=454) and .54 for EnRoute students (N=473).

During academic training the ATC Laboratory Phase, a block test is administered for each subject. A comprehensive knowledge test is also administered to measure the degree to which students have learned the academic portions of ATC subjects taught during laboratory training. This comprehensive phase test is administered prior to giving any operational laboratory control problems. Figure 1, earlier, provides a breakdown of the student's total score for this phase of training is derived. Students

471.

Lab Average *65.0%	**Extra Credit 13.00%	Sixth Problem	Extra Credit Instructor Assessment Problem Errors	2.60% 3.90% 6.50%
	**Instructor Assessment 19.50%	Fifth Problem	Extra Credit Instructor Assessment Problem Errors	2.60% 3.90% 6.50%
	Problem Errors 32.50%	Fourth Problem	Extra Credit Instructor Assessment Problem Errors	2.60% 3.9.. 6.50%
		Third Problem	Extra Credit Instructor Assessment Problem Errors	1.30% 1.95% 3.25%
		Second Problem	Extra Credit Instructor Assessment Problem Errors	1.30% 1.95% 3.25%
		First Problem	Extra Credit Instructor Assessment Problem Errors	1.30% 1.95% 3.25%

Controller Skills Test   25.00%

Comprehensive Phase Test   8.00%

Block Average   2.00%

*The lab average constitutes 65% of the total score for this phase of ATC training.

**On each lab problem the instructor gives a performance rating for that problem that is averaged with the student's problem performance. Since the rating is not allowed to be below 40, essentially the student is given a certain amount of extra credit in the computation of the problem average. Each of the six problems is graded by a different instructor.

Figure 1. Components and weights used in computation of the FAA Academy Laboratory Training Total Score.

472.

must complete this phase of training with a total score of 70 or more to pass the FAA Academy and continue on-the-job training at assigned ATC facilities.

Method

The data were analyzed by multiple regression analysis, using the experimental test battery with and without OKT, to determine the multiple correlation of the test battery with the pass-fail status of the trainees. Separate regression analyses were conducted for the total group, for subgroups of men and women, and for race-ethnic groups where subsample sizes warranted.

The selection utility of the test battery was examined to assess the impact on training costs, and the impact on the sample of ATC trainees in terms of those who would (or would not) have been hired if the test battery had been used for selection decisions. In addition, an analysis was made of the pass-fail status of those who would have been eligible or ineligible for hiring based on the experimental test battery.

RESULTS

Final Sample

As indicated earlier, the total sample of 953 ATC trainees was reduced by missing data on a number of variables. Table 9 shows the number of trainees who had data available on each of the critical variables, by sex and race-ethnic groups. It is clear that the largest reduction of the sample was occasioned by the loss of over 370 cases on the two CSC tests. In order to assess bias due to the missing data, means and standard deviations of the total sample and for those with and without the CSC tests were compared on the two MCAT scores, the OKT, the total CSC battery score, the CSC battery earned rating, and pass-fail. These data are shown in Table 10; note that the sample size varies for the different measures compared.

The t-tests in Table 10 (and the z-test for pass-fail) refer to the differences between the means of the sample with CSC 24 and 157 and the sample without these tests.

It is not surprising that significant differences were observed between the groups on some of the measures, in view of the sample sizes. Their practical significance, however, is a matter of judgment. One appropriate index of practical significance is the estimated effect size, as defined by Cohen (1969). For differences between means, the value of "d" is an estimate of the proportion of the standard deviation represented by the actual difference between the parameters. The value "h" is estimated by the difference between the arcsine values of the two sample proportions. Cohen categorized the effect size of d or h as large (.80), medium (.50), and small (.20). Inspection of the effect sizes in Table 10 shows that all effects were below what Cohen classifies as small, with the exception of the CSC transmuted score, which slightly exceeds the criterion for small.

Table 9

Numbers of trainees with data available, by known status as pre-developmental or non-predevelopmental, sex, and race-ethnic groups, on the new tests (MCAT 1 and 2 and OKT), the CSC tests in the experimental battery (CSC 24 and 157), and pass-fail in training.

Groups	Status as Predevel. or Non-Predevel.	MCAT 1 MCAT 2 OKT	CSC 24 CSC 157	Pass-Fail
Men	788	800	515	790
Women	133	141	67	137
	921	941	582	927
White	821	839	545	821
Black	90	81	39	79
Hispanic	23	23	7	23
Asian	9	10	1	9
	933	953	592	932

474.

Table 10

Comparison of the total sample and those trainees with and without
CSC 24 and 157 on MCAT, OKT, and CSC battery scores (raw and earned
rating) and on pass-fail percentage.

Variable	Total Sample			Sample with CSC 24 & 147			Sample without CSC 24 & 157			Signif. Test t or Z	Estim. Effect Size d
	N	M or p̄	S.D.	N	M or p̄	S.D.	N	M or p̄	S.D.		
MCAT 1	953	36.88	7.12	592	37.07	6.94	361	36.56	7.40	1.07	.07
MCAT 2	953	42.59	5.96	592	43.01	5.87	361	41.90	6.05	2.80**	.19
OKT	953	63.92	15.64	592	63.23	15.82	361	65.04	15.31	-1.75	.12
CSC raw score	704	87.51	7.69	591	87.21	7.62	113	89.08	7.88	-3.10**	.24
CSC E.R.	651	94.93	5.66	540	95.07	5.43	111	94.22	6.61	1.44	.14
Pass/Fail (p̄)	939	.63	.48	585	.62	.49	354	.64	.48	-.62¹⁾	.04

**p < .01

1)z-test; the remainder of entries in this column are t-tests between the means of the sample with
and the sample without CSC 24 and 157.

Further analyses were based on sex and race, in addition to test scores. Test scores on MCAT 1, MCAT 2, and OKT were examined for statistically significant differences by race and sex, between the trainees with scores on CSC 24 and 157 and those without scores on these tests. Table 11 shows the results. No t-tests were computed for the Hispanic and Asian groups because of the small sample sizes. It appears that group differences were not significant for the women or the blacks. Statistically significant differences were found for the men and the whites, for which the sample sizes were large. However, the D values for estimated effect size in Table 11 were all below Cohen's criterion for small.

Whether the two groups can be considered equivalent cannot be answered definitively. However, the analysis supports the assumption that they are, considering that --

No group differences were found on four of the measures (Table 10).

No group differences were found for the women or blacks.

Where group differences were found, the estimated effect sizes were small.

Where group differences were statistically significant, they were in a positive direction for MCAT, but a negative direction for OKT. They did not consistently favor one group or the other.

Table 12 shows the distribution of the total sample of 953 trainees, in comparison to the 592 trainees who had scores on all test variables, by predevelopmental and non-predevelopmental groups.

Test Results

Of the total of 953 trainees in the sample, complete data on all test variables were available for 582, described by sex, and 592, described by race-ethnic group. Total weighted test battery scores, based on MCAT 1 and 2, CSC 157, and CSC 24, were computed for this total reduced sample and for the subsamples of non-redevelopment and predevelopmental trainees, by sex and race-ethnic groups and pass-fail on the test battery. At this point, OKT scores were not considered. The tests were given weights as follows:

Battery Score = 2(MCAT 1)+2(MCAT 2)+1(CSC 157)+1(CSC 24)

Means and standard deviations for the various groups are shown in Table 13. It should be pointed out that the sum of passed and failed trainees in some cases does not equal the number shown for the total group. For example, the sum of the 320 men who passed and the 189 who failed is 509, compared to 515 shown for the total group. This occurred because of the six trainees in the total group for whom information by sex was not available.

Table 11

Comparison of sex and race-ethnic groups with and without
CSC 24 and 157 on MCAT 1 and 2, MCAT Total Score, and OKT

Group	Test	Total Sample			Sample with CSC 24 & 157			Sample without CSC 24 & 157			Signif. Test	Est. Effect Size
		N	Mean	SD	N	Mean	SD	N	Mean	SD	t	d
Men	MCAT 1	800	37.4	6.9	515	37.6	6.6	285	37.0	7.3	2.31*	.09
	MCAT 2	800	42.9	5.8	515	43.3	5.7	285	42.2	6.0	6.11**	.19
	MCAT TOT	800	80.3	11.25	515	80.9	10.9	285	79.2	11.8	2.46*	.15
	OKT	800	64.9	15.5	515	64.1	15.8	285	66.5	14.8	-1.85	.16
Women	MCAT 1	141	34.1	7.8	67	33.3	8.0	74	34.8	7.7	- .81	.19
	MCAT 2	141	40.7	6.4	67	40.7	6.6	74	40.6	6.3	.08	.02
	MCAT TOT	141	74.8	13.1	67	74.0	13.3	74	75.4	13.0	- .27	.11
	OKT	141	57.7	15.1	67	56.3	14.3	74	58.9	15.8	- .38	.17
White	MCAT 1	839	37.6	6.7	545	37.7	6.6	294	37.4	6.9	1.25	.04
	MCAT 2	839	43.2	5.5	545	43.5	5.4	294	42.5	5.6	6.25**	.18
	MCAT TOT	839	80.7	10.8	545	81.2	10.5	294	79.9	11.1	2.17*	.12
	OKT	839	63.9	15.8	545	63.1	15.8	294	65.4	15.5	-1.78	.15
Black	MCAT 1	81	29.8	7.9	39	29.7	7.7	42	29.8	8.2	- .03	.01
	MCAT 2	81	36.6	6.8	39	36.3	7.5	42	36.9	6.1	- .26	.09
	MCAT TOT	81	66.4	13.2	39	66.0	13.9	42	66.7	12.8	- .01	.05
	OKT	81	66.2	13.1	39	65.6	13.4	42	66.8	13.0	- .14	.09
Hispanic	MCAT 1	23	37.1	6.6	7	33.4	6.7	16	38.8	6.1		
	MCAT 2	23	42.7	7.4	7	40.1	5.7	16	43.8	8.0		
	MCAT TOT	23	79.8	12.2	7	73.5	9.5	16	82.6	12.5		
	OKT	23	59.0	20.0	7	60.4	24.7	16	58.4	18.4		
Asian	MCAT 1	10	36.2	4.7	1	31.0	-	9	36.8	4.6		
	MCAT 2	10	43.2	5.7	1	42.0	-	9	43.3	6.0		
	MCAT TOT	10	79.4	9.5	1	73.0	-	9	80.1	9.8		
	OKT	10	61.3	13.7	1	86.0	-	9	58.6	11.2		

Table 12

Composition of the Initial and Reduced Sample for all trainees, predevelopmental trainees, and non-predevelopmental trainees, by sex and race-ethnic groups.

| | All Trainees | | | | Predevelopmental | | | | Non-Predevelopmental | | | |
| | Initial Sample | | Reduced Sample | | Initial Sample | | Reduced Sample | | Initial Sample | | Reduced Sample | |
Group	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Men	800	85.0	515	88.5	61	55.5	20	51.3	739	88.9	495	91.2
Women	141	15.0	67	11.5	49	44.5	19	48.7	92	11.1	48	8.8
Total	941	100.0	582	100.0	110	100.0	39	100.0	831	100.0	543	100.0
White	839	88.0	545	92.0	42	38.1	13	32.5	797	94.5	532	96.4
Black	81	8.5	39	6.6	55	50.0	24	60.0	26	3.1	15	2.7
Hispanic	23	2.4	7	1.2	12	10.9	3	7.5	11	1.3	4	.7
Asian	10	1.1	1	.2	1	1.0	-	-	9	1.1	1	.2
Total	953	100.0	592	100.0	110	100.0	40	100.0	843	100.0	552	100.0

478.

Table 13

Means and standard deviations for sex and race-ethnic groups
on the weighted experimental battery scores in the total reduced
sample and the non-predevelopmental and predevelopmental sub-
samples who passed and failed the battery.

Total Trainees

Group	Pass N	Pass Mean	Pass SD	Fail N	Fail Mean	Fail SD	Total N	Total Mean	Total SD
Men	320	255.0	22.5	189	234.2	28.9	515	247.2	26.9
Women	34	246.6	22.9	32	223.2	32.4	67	235.1	29.9
White	345	254.7	22.7	194	237.0	27.6	545	248.3	25.9
Black	12	240.0	15.9	26	198.5	23.1	39	211.5	28.2
Hispanic	4	244.8	27.3	3	221.0	12.5	7	234.6	24.2
Asian	1	233.0	-	-	-	-	1	233.0	-
Total	-	-	-	-	-	-	592	245.6	27.6

Non-predevelopmental

Group	Pass N	Pass Mean	Pass SD	Fail N	Fail Mean	Fail SD	Total N	Total Mean	Total SD
Men	313	255.3	22.4	176	236.7	27.4	495	248.5	25.8
Women	26	248.8	22.6	21	233.3	30.5	48	241.4	27.1
White	338	254.9	22.6	188	237.6	27.4	532	248.3	25.9
Black	5	236.2	15.4	9	203.0	18.7	15	214.3	22.9
Hispanic	2	267.5	10.6	2	228.0	4.2	4	247.8	23.7
Asian	1	233.0	-	-	-	-	1	233.0	-
Total	-	-	-	-	-	-	552	247.6	26.1

Predevelopmental

Group	Pass N	Pass Mean	Pass SD	Fail N	Fail Mean	Fail SD	Total N	Total Mean	Total SD
Men	7	242.4	24.8	13	200.4	27.2	20	215.1	33.0
Women	8	239.5	23.9	11	204.0	27.8	19	218.9	31.3
White	7	248.7	30.9	6	217.8	29.3	13	234.5	33.0
Black	7	242.7	16.9	17	196.2	25.4	24	209.7	31.5
Hispanic	2	222.0	7.1	1	207.0	-	3	217.0	10.0
Asian	-	-	-	-	-	-	-	-	-
Total	-	-	-	-	-	-	60	218.3	32.5

The same weighting formula was also applied, with the addition of OKT, to evaluate the added influence of that test. OKT was not intended to be included in the decision concerning pass-fail on the test battery, but rather as an alternative to the OPM Rating Guide, as used in 1978, to award extra credit for aviation-related knowledge. In this procedure, OKT was given a weight of 1:

Battery Score = 1(MCAT 1)+2(MCAT 2)+1(CSC 157)+1(CSC 24)+1(OKT)

Means and standard deviations for these scores are shown in Table 14. Here it is apparent that the differences in test performance of the groups that passed and failed the battery were consistent. For example, the mean scores of all the pass groups were above 300 and of all the fail groups, below 300, and the differences between pass and fail group means were generally around one standard deviation. This pattern also held generally for the weighted test battery with OKT scores excluded (Table 13).

Intercorrelations and Multiple Regression Analysis

Table 15 presents the intercorrelations of the 2 MCAT forms, the OKT, and CSC tests 157 and 24 and the correlations of each with pass-fail in the Academy, for the total reduced sample and for men, women, whites, and blacks; other race-ethnic groups were excluded because of small sample size. The correlations with pass-fail and zero-order validity coefficients for the restricted trainee population. Correction for restriction of range was not a concern in this analysis, since the objective was to compute regression equations and determine the relative contribution of the component tests to the prediction of the criterion (pass-fail in training). However, in order to afford some understanding of the effect of restriction of range in this case, Table 16 compares the restricted validity coefficients reported by Boone (1979b) for the same tests on the 1976-1978 sample of 1827 ATC trainees. It is apparent that the estimated true coefficients were substantially higher than the restricted correlations.

The restricted correlations were included in the calculation of multiple regression equations, as shown in Table 17. The OKT was excluded from these calculations. The table shows the predictors in a constant order, with R (the multiple correlation coefficient) and R^2 incremented as each variable was entered. Note that the Beta weights vary in the sex and race-ethnic group samples. Of the four tests, CSC 24 and 157 made the largest contributions to prediction for the women and blacks, although their contributions in the total sample were small. This is emphasized in the following tabulation:
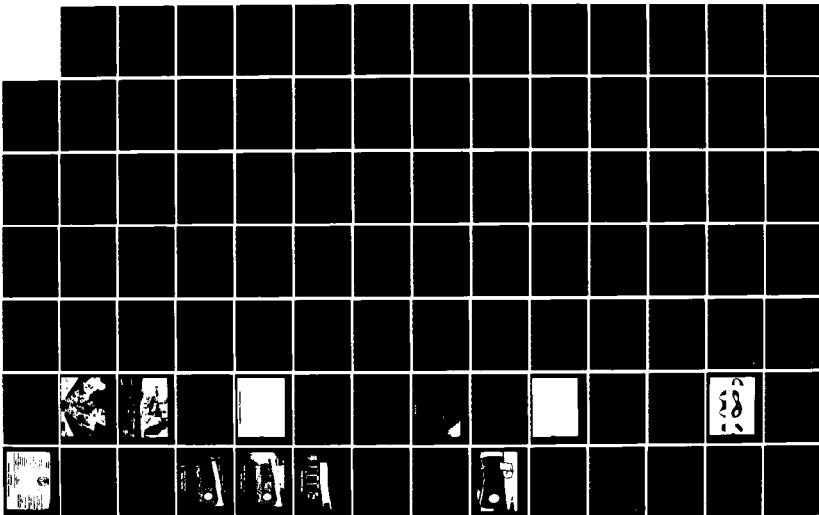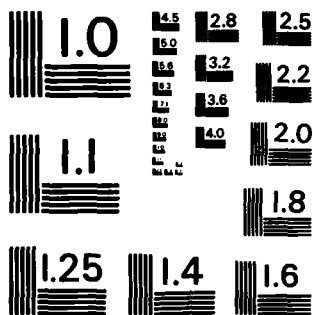
MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

Table 14

Means and standard deviations for sex and race-ethnic groups on the weighted experimental battery scores (including OKT weighted 1) in the total reduced sample and the non-predevelopmental and predevelopmental subsamples who passed and failed the battery.

Total Trainees

Group	Pass N	Pass Mean	Pass SD	Fail N	Fail Mean	Fail SD	Total N	Total Mean	Total SD
Men	320	321.7	25.6	189	294.0	29.8	515	311.3	30.3
Women	34	307.3	28.5	32	275.2	33.3	67	29.14	34.5
White	345	320.5	26.3	194	295.2	29.4	545	311.3	30.0
Black	12	310.2	18.0	26	261.8	28.7	39	277.1	33.8
Hispanic	4	312.0	34.3	3	272.3	11.0	7	295.0	32.8
Asian	1	319.0	-	-	-	-	1	319.0	-
Total	-	-	-	-	-	-	592	309.0	31.4

Non-predevelopmental

Group	Pass N	Pass Mean	Pass SD	Fail N	Fail Mean	Fail SD	Total N	Total Mean	Total SD
Men	313	321.6	25.6	176	295.9	28.6	495	312.2	29.4
Women	26	308.9	29.0	21	278.4	34.4	48	294.7	34.6
White	388	320.6	26.0	188	295.6	29.4	532	311.5	29.8
Black	5	302.0	8.2	9	259.9	18.2	15	275.0	24.8
Hispanic	2	321.0	55.2	2	272.0	15.6	4	296.5	43.5
Asian	1	319.0	-	-	-	-	1	319.0	-
Total	-	-	-	-	-	-	552	310.5	30.3

Predevelopmental

Group	Pass N	Pass Mean	Pass SD	Fail N	Fail Mean	Fail SD	Total N	Total Mean	Total SD
Men	7	324.4	27.8	13	268.7	34.5	20	288.2	41.7
Women	8	302.3	28.1	11	269.0	31.6	19	283.0	33.9
White	7	315.4	39.6	6	285.2	29.0	13	301.5	37.1
Black	7	316.0	21.3	17	262.9	33.4	24	278.4	38.9
Hispanic	2	303.0	12.7	1	273.0	-	3	293.0	19.5
Asian	-	-	-	-	-	-	-	-	-
Total	-	-	-	-	-.	-	40	287.0	38.1

Table 15

Intercorrelations and correlations with pass-fail in Academy
training of the MCAT 1, MCAT 2, OKT, CSC 157, and CSC 24, for
the total reduced sample, men, women, white, and black trainees.

Total Reduced Sample

(Minimum N=585 -- Criterion and CSC 24)

Test	MCAT 1 (N=953)	MCAT 2 (N=953)	OKT (N=953)	CSC 24 (N=592)	CSC 157 (N=592)	Pass-Fail (N=585)
MCAT 1	-	.60	.08	.29	.21	.35
MCAT 2		-	.12	.20	.20	.37
OKT			-	-.24	-.14	.25
CSC 24				-	.10	.10
CSC 157					-	.18

MEN

(Minimum N=509 -- Criterion and CSC 24)

Test	MCAT 1 (N=800)	MCAT 2 (N=800)	OKT (N=800)	CSC 24 (N=515)	CSC 157 (N=515)	Pass-Fail (N=509)
MCAT 1	-	.57	.05	.32	.22	.34
MCAT 2		-	.08	.25	.21	.37
OKT			-	-.23	-.15	.21
CSC 24				-	.10	.12
CSC 157					-	.16

WOMEN

(Minimum N=66 -- Criterion and CSC 24)

Test	MCAT 1 (N=141)	MCAT 2 (N=141)	OKT (N=141)	CSC 24 (N=67)	CSC 157 (N=67)	Pass-Fail (N=66)
MCAT 1	-	.69	.11	.12	.22	.37
MCAT 2		-	.17	-.07	.19	.38
OKT			-	-.25	.07	.31
CSC 24				-	.13	.08
CSC 157					-	.40

WHITE

(Minimum N=539 -- Criterion and CSC 24)

Test	MCAT 1 (N=839)	MCAT 2 (N=839)	OKT (N=839)	CSC 24 (N=545)	CSC 157 (N=545)	Pass-Fail (N=539)
MCAT 1	-	.55	.10	.28	.21	.29
MCAT 2		-	.13	.19	.20	.33
OKT			-	-.26	-.13	.23
CSC 24				-	.10	.07
CSC 157					-	.17

BLACK

(Minimum N=38 -- Criterion and CSC 24)

Test	MCAT 1 (N=81)	MCAT 2 (N=81)	OKT (N=81)	CSC 24 (N=39)	CSC 157 (N=39)	Pass-Fail (N=38)
MCAT 1	-	.63	.33	.15	-.29	.56
MCAT 2		-	.22	.06	-.24	.45
OKT			-	-.10	-.12	.24
CSC 24				-	-.07	.35
CSC 157					-	.07

Table 16

Comparison of restricted validity coefficients in the present study
with restricted and corrected coefficients for the same tests
reported by Boone (1979b) in an earlier study.

| | ATC Trainees* Jan. 1976 - April 1978 | | ATC Trainees** June 1978 - Dec. 1978 |
	Restricted r	Corrected r	Restricted r
MCAT Exp. (Z scores)	.28	.53	-
MCAT 1 (Raw scores)	-	-	.35
MCAT 2 (Raw scores)	-	-	.37
CSC 24	.10	.34	.10
CSC 157	.07	.40	.18
OKT	.22	-	.25

*Criterion = ATC Lab. Average Z scores (See Chapter 18, Part III,
Tables 17 and 18)
**Criterion = Academy Training Pass-Fail (Lab. Training)

Table 17

Step-wise multiple correlation (R) and multiple correlation squared (R^2), and beta weights for tests in the experimental battery, for reduced total sample and subsamples of men, women, whites, and blacks.

Test

Group	MCAT 1			MCAT 2			CSC 24			CSC 157		
	R	R^2	beta	R	R^2	beta	R	R^2	beta	R	R^2	beta
Total (N=585) Sample	.348	.121	.1832	.405	.164	.2470	.405	.164	-.0109	.416	.173	.0997
Men (N=509)	.337	.114	.1800	.400	.160	.2554	.400	.160	-.0121	.406	.165	.0691
Women (N=66)	.375	.140	.1430	.412	.170	.2259	.419	.175	.0335	.523	.273	.3240
White (N=539)	.292	.085	.1538	.355	.126	.2294	.355	.126	-.0270	.368	.135	.0982
Black (N=38)	.565	.319	.4734	.579	.335	.2016	.640	.409	.2836	.689	.475	.2697

484.

Increment to R over preceding tests in regression equation:

Group	CSC 24	CSC 157
Men	-	.006
Women	.007	.104
Black	.061	.049
Total sample	-	.011

Other regression analyses were computed for the total group, with CSC 24 excluded; in these, the multiple correlations and beta weights obtained were essentially unchanged. For the total sample, the results for CSC 24 were consistent with those for previous trainee samples and applicant samples. In the Boone (1979b, See Chapter 18, Part III) study of 1827 1976-1978 trainees, CSC 24 contributed least to the multiple correlation and had the smallest beta weight. In the 1976-1977 applicant sample (Rock et al., 1982, Chapter 8), CSC 24 contributed least to passing scores on the CSC battery. However, since CSC 24 and 157 did contribute to the prediction of the criterion for some of the groups, it was decided to include them in the battery. Thus the weighted test battery remained:

$$2(MCAT\ 1)+2(MCAT\ 2)+1(CSC\ 157)+1(CSC\ 24).$$

Table 18 shows the step-wise multiple regressions for the battery with OKT included. In these analyses, the order in which each test entered each regression analysis was determined by its contributions to the multiple correlation coefficient. For each group, the tests entered in a different order; the order for women and blacks differed most from that of the total group.

Table 19 presents in summary form the multiple correlation coefficients (R) and the percentage of criterion variance (R^2) accounted for by the respective prediction equations that were based on restricted zero-order correlation coefficients. These are presented for the total reduced sample, whites, men, women, and blacks, for the experimental test battery excluding OKT and including OKT.

Test Utility

An analysis was carried out to examine the practical utility of the new ATC test battery, to relate the findings to the information obtained on the 1978 ATC Applicant Group, and to estimate the potential savings in FAA training investment that would be realized by using the new ATC test for selection of applicants.

As noted previously, the weighted test scores for this sample of 585 ATC trainees was based on:

$$WT_1 = 2(MCAT\ 1)+2(MCAT\ 2)+1(CSC\ 157)+1(CSC\ 24)$$

485.

Table 18

Step-wise multiple correlation (R), multiple correlation squared (R^2), and beta weights for tests in the experimental battery, including OKT, for the reduced total sample and subsamples of men, women, whites, and blacks.

Group	N	Test	R	R^2	beta
Total Sample	585	MCAT 2	.374	.140	.2134
		OKT	.429	.184	.2478
		MCAT 1	.455	.207	.1547
		CSC 157	.474	.225	.1398
		CSC 24	.477	.228	.0588
Men	509	MCAT 2	.369	.136	.2264
		OKT	.421	.177	.2341
		MCAT 1	.448	.200	.1539
		CSC 157	.460	.211	.1109
		CSC 24	.462	.214	.0530
Women	66	CSC 157	.403	.162	.2997
		OKT	.531	.282	.3352
		MCAT 1	.592	.351	.1294
		MCAT 2	.603	.363	.1873
		CSC 24	.613	.376	.1199
White	539	MCAT 2	.328	.108	.1924
		OKT	.398	.158	.2583
		CSC 157	.424	.180	.1397
		MCAT 1	.438	.192	.1193
		CSC 24	.440	.194	.0476
Black	38	MCAT 1	.565	.319	.3573
		CSC 24	.625	.391	.3360
		OKT	.698	.488	.3415
		CSC 157	.744	.553	.2791
		MCAT 2	.759	.576	.1974

Table 19

R and R^2 for prediction of pass-fail in training
for reduced sample of 1978 trainees, for experimental
test battery excluding and including OKT, for total
sample, white, men, women, and blacks.

| | | Test Battery | | | |
| | | Excluding OKT | | Including OKT | |
Group	N	R	R^2	R	R^2
Total Sample	585	.416	.173	.477	.228
Whites	539	.368	.135	.440	.194
Men	509	.406	.165	.462	.214
Women	66	.523	.273	.613	.376
Blacks	38	.689	.475	.759	.579

However, for the 1978 ATC Applicant Group, only one parallel form of MCAT was administered rather than the two which were given to the present sample of ATC trainees. In order to relate the two data sets, the weighted ATC test scores for the 585 ATC trainees were recomputed using only the first MCAT administered and then weighting each of the tests on the same basis as in the analysis of the 1978 ATC Applicant Group. The weighting used was:

$$WT_2 = 4(MCAT\ 1) + 2(CSC\ 157) + 1(CSC\ 24)$$

The total raw weighted scores were then transmuted to a scale of 0-100 for each trainee, based on:

$$T_s = 10.68(\frac{RWS - 222.27}{56.22}) = 70$$

as was shown with Table 28, Chapter 20. Again, a transmuted score of 70 or above was defined as the passing score for the new ATC test battery for this ATC trainee group.

In order to assess the difference between these two weighting methods (WT_1 and WT_2), correlation coefficients were computed between the two weighted scores on the experimental battery and between those and the scores on the CSC battery which were also available for the ATC trainees. Table 20 presents the descriptive statistics and correlations.

Since the weighting used to examine the utility of the experimental test battery (WT_2) had a slightly higher correlation with the CSC battery than the battery (WT_1) (.65 compared to .62), the results should tend to be somewhat more conservative than if the analysis were based on weighting factors identified in WT_1.

The mean transmuted score on the CSC battery for this group of ATC trainees was 87.2 with a standard deviation of 7.6. Their mean transmuted score on the experimental battery (WT_2) was 79.6 with a standard deviation of 6.7. This lower mean score is consistent with the findings derived from the 1978 ATC Applicant Group (Chapter 20), which demonstrated that applicants who passed both the CSC battery and the experimental battery generally scored lower on the experimental battery.

For the present sample of 585 ATC trainees, Table 21 compares the score distribution for the CSC battery and the experimental battery for the total sample and for the predevelopmental and non-predevelopmental groups. These are transmuted test scores without veterans preference or aviation-related knowledge/experience credit. The 585 ATC trainees for whom complete test and criterion data were available represent 61 percent of the 953 attending the Academy during June through December 1978. The 40 predevelopmental trainees represent 36 percent of the non-competitive hires; the 545 non-predevelopmental trainees represent 65 percent of the competitive hires.

488.

Table 20

Correlations of weighted experimental battery (Wt. 1 and Wt. 2) and CSC battery for 1978 trainee reduced sample.

Test Battery	N	Mean	S.D.	Correlations 12	13	23
1. Exper. Battery Wt. 1	592	245.6	27.57	.93	.62	.65
2. Exper. Battery Wt. 2	592	272.6	35.10			
3. CSC Battery	591	262.0	23.31			

Table 21

Distribution of transmuted scores of the 1978 trainee reduced sample on the CSC Battery and the Experimental Battery (Wt. 2), for total sample, non-predevelopmental, and predevelopmental trainees.

Group	Score Range	CSC Battery N	CSC Battery Percent	Experimental Battery N	Experimental Battery Percent
Total sample N=585	95+	121	20.7	2	.3
	90-94	125	21.4	27	4.6
	85-89	133	22.6	98	16.8
	80-84	110	18.8	161	27.6
	75-79	63	10.8	147	25.1
	70-74	33	5.7	102	17.4
	Subtotal - Pass	585	100.0	537	91.8
	65-69	0	0	39)	
	60-64	0	0	8)	8.2
	55-59	0	0	1)	
	Subtotal - Fail	0	0	48	
	Mean	87.2		79.6	
	S.D.	7.6		6.7	
Non-predevel. N=545	95+	119	21.8	2	.4
	90-94	122	22.4	27	5.0
	85-89	129	23.7	97	17.8
	80-84	101	18.5	157	28.8
	75-79	48	8.8	136	24.9
	70-74	26	4.8	90	16.5
	Subtotal - Pass	545	100.0	509	93.4
	65-69	0	0	31)	
	60-64	0	0	5)	6.6
	55-59	0	0	0)	
	Subtotal - Fail	0	0	36	
Predevel. N=40	95+	2	4.6	0	0
	90-94	3	7.0	0	0
	85-90	4	9.3	1	2.5
	80-84	9	23.3	4	10.0
	75-79	15	37.2	11	27.5
	70-74	7	18.6	12	30.0
	Subtotal - Pass	40	100.0	28	70.0
	65-69	0	0	8)	
	60-64	0	0	3)	
	55-59	0	0	1)	30.0
	Subtotal - Fail	0	0	12	

Two significant points are evident from the data in Table 21. First, 48 (8.2%) of the 585 ATC trainees, all of whom passed the CSC battery for employment eligibility, would not have passed the experimental battery and therefore would not have been eligible for appointment as air traffic controller trainees. When the total group is further identified by competitive hires from established OPM certificates (non-predevelopment and non-competitive hires (predevelopmental) who were hired through alternate recruitment procedures, other differences are apparent: (1) 30% of the predevelopmental group would not have been eligible for appointment, in contrast to 6.6% of the non-predevelopment hires (this 6.6% compares to 7.2%) of the 1978 ATC Applicant Group who also passed the CSC battery but did not pass the weighted experimental battery (WT_2); and (2) 23% of the non-predevelopmental trainees scored 85 or above on the experimental battery, in contrast to 2.5% of the predevelopmental group. These data provide some further quantitative insight into the mean score differences between these two groups which were identified in the preceding discussion.

Second, the score distribution of these ATC trainees on the experimental battery was markedly lower than on the CSC battery. About 42% of the total trainee group scored 90 or above on the CSC battery in contrast to only 5% on the experimental battery. In the non-predevelopmental group, 44% scored 90 or higher on the CSC battery, compared to 5.4% on the experimental battery. None of the predevelopmental trainees scored above 89 on the experimental battery, while almost 12% did on the CSC battery.

These findings are consistent with the results obtained for the 1978 ATC applicants, for whom passing scores on the experimental battery were significantly lower than on the CSC battery.

Given this marked difference in score distribution between the CSC battery and the experimental battery, how were the scores on both batteries related to pass-fail in training? Test scores and criterion measures were available for 695 trainees for the CSC battery and for 585 trainees for the experimental battery (WT_2). The results are shown in Table 22. Again, these scores do not include extra credit for veterans preference or aviation-related experience or knowledge. This table clearly identifies the difference between the CSC battery and the experimental battery with regard to the percentages of trainees who failed the Academy training program in each score range. For example, in the score range 90-94, almost 32% on the CSC battery and only 19% on the experimental battery failed. It is noteworthy that of the 48 trainees who passed the CSC battery and scored below 70 on the experimental battery, 79% failed in training.

Table 23 shows pass-fail data for non-predevelopmental and predevelopmental trainees in relation to the experimental battery. The difference in failure rate between the two groups (36.5% vs 60.0%) is striking. In addition, the 48 trainees who scored below 70 (and would have been ineligible for employment if the experimental battery had been operational in place of the CSC battery) were divided 3/4 (36) in the non-predevelopmental group and 1/4 (12) in the developmental group, and their failure

491.

Table 22

Distributions and fail rates in training of 1978 trainees, in
relation to raw scores on the CSC Battery and the Experimental Battery.

Score Range	CSC Battery				Experimental Battery			
	Total N	Sample Percent	Failed N	Training Percent	Total N	Sample Percent	Failed N	Training Percent
95+	145	20.0	33	22.8	2	.3	0	0
90-94	148	21.3	47	31.8	27	4.6	5	18.5
85-89	156	22.5	62	39.7	98	16.8	20	20.4
80-84	131	18.8	59	45.0	164	27.6	46	28.0
75-79	76	10.9	38	50.0	147	25.1	62	42.2
70-74	39	5.6	22	56.4	102	17.4	52	51.0
Below 70	0	0	0	0	48	8.2	38	79.2
Total	695	100.0	261	37.6	585	100.0	223	38.1

Table 23

Distributions and fail rates of non-predevelopmental and predevelopmental 1978 trainees in relation to scores on the experimental battery.

| Score Range | Non-predevelopmental | | | | Predevelopmental | | | |
| | N | Failed Training | | | N | Failed Training | | |
		N	Pct. of Tot.	Rate		N	Pct. of Tot.	Rate
95+	2	0	0	0	0	0	0	0
90-94	27	5	2.5	18.5	0	0	0	0
85-89	97	20	10.1	20.6	1	0	0	0
80-84	157	46	23.1	29.3	4	0	0	0
75-79	136	57	28.6	41.9	11	5	20.8	45.5
70-74	90	43	21.6	47.8	12	9	37.5	75.0
Below 70	36	28	14.1	77.8	12	10	41.7	83.3
Total	545	199	100.0	36.5	40	24	100.0	60.0

493.

rates were 77.8% and 83.3%, respectively. Had these 48 trainees not been hired, the overall loss rate would have been reduced from 38.1% to 34.4% (and for the non-predevelopmental and predevelopmental groups, from 36.5% to 33.6% and from 60% to 50%, respectively.).

In 1978, for each predevelopment trainee, FAA invested about $28,000 in salary and training costs and about $10,000 in each non-predevelopmental student by the time their FAA Academy ATC training program was completed. For each student who failed or withdrew from training, this investment was lost and a new trainee had to be hired. If the 48 trainees who scored below 70 on the experimental battery had not been hired, FAA's loss in training investment costs would been reduced by $560,000. Extending this for 2,000 new trainees entering Academy training, or an annual basis this would have resulted in a cost avoidance of about $1,900,000 each year in 1978 dollars.

This projection of the cost avoidance utility of the experimental test battery addresses only the benefit obtained if ATC applicants who scored below a minimum of 70 were not hired. The other significant area where benefits could be realized involves the differences between the applicants who would be hired under competitive selection procedures with the CSC battery and the experimental battery, after allowance of extra credit for veterans preference and aviation-related experience or knowledge.

In order to examine these differences, an analysis was made using OPM Earned Rating, which was derived from the trainee's score on the CSC battery plus additional credit for veterans preference (0 to 10 points) and credit for aviation-related experience provided by the OPM Rating Guide (0-15 points). This was available for 621 of the 843 non-predevelopmental ATC trainees in this sample. The 110 predevelopmental ATC trainees, who were appointed non-competitively, were not included in this analysis since extra credit for veterans preference or aviation-related experience was not a factor in their appointment eligibility; they needed only to pass the CSC battery with a score of 70 or above. An "equivalent" of the OPM Earned Rating was computed for the new experimental battery (WT$_2$), using the OKT scores. The 36 trainees who did not achieve a minimum score of 70 on the experimental battery were excluded from this analysis, since they would not have been eligible for appointment. The Earned Rating for these trainees was based on their scores on the experimental battery plus additional credit provided for their OKT scores as follows:

OKT Scores	Additional Points
80+	15
75-79	10
70-74	5
65-69	3

494.

The difference between the CSC transmuted score and the OPM Earned Rating for each trainee was the additional credit given for veterans preference and aviation experience based on the OPM Rating Guide. Since the veterans preference points were not separately identified in the total of extra credit given in the OPM Earned Rating, it was not possible specifically to identify those ATC trainees who were 5-point or 10-point veterans. In order to derive an "equivalent" Earned Rating on the experimental battery that would reflect extra credit for veterans preference and aviation-related knowledge based on OKT scores, a set of rules was followed, which attempted to estimate V_1, which was defined as the total extra credit received (the OPM Earned Rating minus the CSC battery transmuted score). For the total sample, the mean Earned Rating was 95.1 (N=540) and the CSC transmuted score was 87.2 (N=591). The mean V_1 was 7.9 points.

The rules developed were the following if-then statements:

IF:	THEN
V_1 is less than 5	VET points = 0
	OKT credit = 0, 3, 5, 10, 15
$V_1 = 5$	VET points = 5
	OKT credit = 0, 3, 5, 10, 15
$V_1 = 10, 15,$ or 20	VET points = 5
	OKT credit = 0, 3, 5, 10, 15
V_1 is more than 20	VET points = 10
	OKT credit = 0, 3, 5, 10, 15

It was expected that the net effect of this procedure would be a somewhat higher distribution of Earned Rating scores for the experimental battery than would have been the case if the actual veterans credit had been available for every trainee. Following the rules outlined, an estimated total Earned Rating could be computed based on the experimental battery for 486 of the ATC trainees appointed competitively (non-predevelopmental) from OPM registers. The 48 students who scored below 70 on the experimental battery were excluded from the sample. Table 24 compares the distribution of these Earned Ratings with the Earned Ratings for 621 trainees based on the CSC battery and Rating Guide. The 486 trainees for whom estimated Earned Ratings could be computed using the experimental battery represent a subsample of the 621 trainees who had OPM Earned Ratings available.

About 84% of the 621 trainees had OPM Earned Ratings of 90 or more. Since competitive selection from OPM registers requires hiring those applicants with the highest Earned Ratings first, this distribution of OPM Earned Ratings was expected. Nevertheless, the Earned Rating distribution based on the Experimental battery for the 486 trainees who were part of the sample of 621, is quite different. Only 41% of the trainees had Earned

Table 24

Comparison of Distributions of OPM Earned Ratings
(based on CSC Battery) and Experimental Battery
estimated Earned Ratings.

Earned Rating Score Range	OPM Earned Rating		Est. Exper. Battery Earned Rating	
	N	Percent	N	Percent
100+	205	33.0	54	11.1
90-99	314	50.6	143	29.4
80-89	99	15.9	229	47.1
70-79	3	.5	60	12.6
Total	621	100.0	486	100.0

Ratings of 90 or above and almost 13% were below 80, compared with less than 1% for the OPM Earned Rating. Consequently, many of the trainees with Earned Ratings below 90 on the experimental battery would not have been hired competitively, since other applicants with higher ratings would have been selected in their place.

In relation to the practical utility of the experimental battery and OKT, these differences in Earned Rating score distribution would be of significance only if the pass-fail rates for the two batteries were differentially favorable to the experimental battery. Data relevant to this issue are shown in Table 25, in which it is clear that the probability of successful completion of the ATC training program was greater in each of the Earned Rating score ranges above 79 for the experimental battery than for the CSC battery.

As a consequence of the foregoing analysis, it is reasonable to expect that use of the Earned Rating on the experimental battery would result in hiring a somewhat different and superior group of ATC trainees than was in fact hired with the 1978 OPM register. It is possible to relate the data in Tables 24 and 25 to the results obtained for the 1978 ATC Applicant Group on both the CSC battery and the experimental battery. This information would provide a basis for assessing the differences between the two groups of applicants selected and the potential effect of the two bases of selection on fail rates and on training investment losses.

The analysis in Chapter 20 of the 1978 ATC Applicant Group (Table 27, Chapter 20) identified the Earned Rating score distribution of applicants who passed both the CSC battery and the weighted experimental battery (WT$_2$). If this group of 621 ATC trainees had been hired from the group of 1978 ATC applicants, based on Earned Ratings obtained on the CSC battery, the relationship would have been as shown in Table 26. For the 1978 ATC Applicant Group, the number in each Earned Rating category in effect represents the OPM register from which applicant selections would have been made.

The registers for ATC applicants are maintained by OPM by separate employing jurisdictions. Consequently, each of the 11 FAA regions selects from its own certificate of applicants, based on the geographic preference identified by the applicant and the Earned Rating. As a result, one region (Southern, for example) might be able to fill all of its ATC recruitment needs with applicants who have Earned Ratings above 100, while other regions (Alaska or Great Lakes, for example) might recruit a number of applicants with Earned Ratings in the 80's or even 70's. Nevertheless, Table 26 shows that essentially all of the 621 ATC trainees scored 80 or above, and the mean OPM Earned Rating was 94.9 for the total group. As a result, selections were made in most cases from the top range of the registers, with scores of 90 or more. The selection ratio in the right-hand column of Table 26 shows the percentage of 1978 applicants who were hired and became 1978 trainees. For example, in the Earned Rating score

Table 25

Distribution of Earned Ratings based on the CSC Battery
(OPM Earned Rating) and estimated for the Experimental
Battery, in relation to pass-fail in training.

Earned Rating Score Range	OPM Earned Rating			Est. Exper. Battery Earned Rating			Percent Diff.
	Total N	Fail/WD Rate N	Percent	Total N	Fail/WD Rate N	Percent	
100+	205	54	26.3	54	6	11.1	-15.2
90-99	314	122	38.9	143	33	23.1	-15.8
80-89	99	47	47.5	229	86	37.6	- 9.9
70-79	3	2	66.6	60	38	66.3	- .3
Total	621	225	36.2	486	163	33.5	

Table 26

.OPM Earned Ratings for qualified 1978 applicants (N=3035) and for 1978 trainees (competitive hires, N=621), based on the Experimental Battery, and Selection Ratio (percent of applicants who would have been hired) at each score range.

OPM Earned Rating Score Range	OPM Earned Rating Qualified 1978 ATC Applicants		OPM Earned Rating 1978 ATC Trainees (Competitive Hires)		Selection Ratio
	N	Percent	N	Percent	Percent
100+	229	7.5	205	33.0	89.5
90-99	644	21.2	314	50.6	48.8
80-89	1086	35.8	99	15.9	9.1
70-79	1076	35.5	3	.5	.3
Total	3035	100.0	621	100.0	20.5

range of 100+ there were 229 applicants and 205 of them became trainees, giving a selection ratio of 89.5%. As a matter of fact, this selection ratio (for the 100+ range) was somewhat higher than that typically experienced by the FAA regions. Generally, of those applicants offered ATC positions, about 80% were selected; the other 20% declined or were dropped from consideration because of medical or security reasons.

Next, the composition of a group of 621 trainees was projected by Earned Rating scores, on the assumption that they had been selected from a competitive register based on the experimental battery (WT_2) including extra credit for OKT and veterans preference. The number of 1978 applicants in each Earned Rating score range, based on the scores, was used as a selection register, as shown in Table 27. This is the same distribution as that presented in Table 27, Chapter 20.

A maximum selection ratio of 80% in each Earned Rating group was used as more representative of actual selection experience. Since fewer applicants were expected to achieve Earned Ratings of 90 or more, based on the experimental battery, more selections would be made from among applicants with Earned Ratings in the 80-99 range. Using these factors, the projected distribution of the 621 trainees, based on the experimental battery, would be as indicated in Table 27. Then, applying the Earned Rating score distribution for this group of 621 trainees based on the CSC battery, and the distribution derived in Table 27, based on the experimental battery, the pass-fail rates shown in Table 25 were applied to the two distributions to estimate failure rates based on the two test batteries. These results are shown in Table 28, which demonstrates the potential utility of the experimental battery in terms of an overall reduction of the fail rate of about 10% (from 36.2% to 26.1%) for this sample of 621 trainees.

The cost-benefit of this reduction in training failures can be estimated by projecting the number of failures that would be avoided on an annual basis for 1800 ATC applicants hired through competitive selection procedures. In 1978 dollars, this would result in a cost avoidance of 1.825 million dollars. By combining this estimate with the 2.350 million reduction in investment loss realized by not hiring those applicants who would have passed the CSC battery but failed the experimental battery is estimated to be around 3 million dollars per year.

Table 27

Estimated composition of a group of 621 1978 trainees based on the application of assumed selection ratios to the distribution of 1978 applicants, by Earned Rating score range on the Experimental Battery.

Exper. Battery Earned Rating Score Range	Qualified 1978 ATC Applicants		Est. 1978 ATC Trainees		Assumed Selection Ratio
	N	Percent	N	Percent	Percent
100+	126	4.0	100	16.1	80.
90-99	402	12.9	322	51.9	80.
80-89	1162	37.3	192	30.9	17.
70-79	1426	45.8	7	1.1	-
Total	3116	100.0	621	100.0	20.

501.

Table 28

Distribution of 1978 trainees and fail rates based on the CSC battery, at left, and on estimated register based on the experimental battery, at right.

	CSC Battery				Experimental Battery			
OPM Earned Rating Score Range	Actual Trainee Distribution N	Percent	Actual Fail Rate N	Percent	Est. Earned Rating Score Range	Est. Trainee Distribution N	Percent	Est. Fail Rate N	Percent
100+	205	33.0	54	26.3	100+	100	16.1	11	11.1
90-99	314	50.6	122	38.9	90-99	322	51.9	74	23.1
80-89	99	15.9	47	47.5	80-89	192	30.9	72	37.6
70-79	3	.5	2	66.6	70-79	7	1.1	5	66.6
Total	621	100.0	225	36.2	Total	621	100.0	162	26.1

Chapter 22

CONFORMITY OF THE NEW EXPERIMENTAL TEST BATTERY TO THE
UNIFORM GUIDELINES ON EMPLOYEE SELECTION REQUIREMENTS [1]

Donald B. Rock, John T. Dailey, Herbert Ozur,
James O. Boone, and Evan W. Pickrel

.

The Uniform Guidelines (1978) established jointly by the Equal Employment Opportunity Commission (EEOC), Department of Labor (DOL), Department of Justice (DOJ), and the Office of Personnel Management (OPM), provides four basic requirements for selection procedures in relation to equal employment opportunity. Briefly, these requirements are:

1. Adverse impact. It is required that a determination be made as whether or not a selection procedure has an adverse impact on employment opportunities of minorities or women;

2. Validation. If there is an adverse impact, then it is necessary to demonstrate the validity of the selection procedure;

3. Fairness. When empirical data demonstrate validity of the selection procedure (that is, it is predictive of or significantly correlated with important elements of job performance), it is necessary to justify the fairness of the selection procedure;

4. Alternative procedures. Consideration and investigation of alternative selection procedures are required. When two or more selection procedures which are substantially equally valid for a given purpose, the procedure which has the lesser adverse impact should be used.

This chapter reviews the research relevant to the Civil Service Commission (CSC) test battery and the experimental test battery and the component tests presented in the preceding chapters of Part IV, in relation to the requirements of the Uniform Guidelines and other elements of EEO programs.

Adverse Impact

For enforcement proceedings, the Uniform Guidelines state:

"A selection rate for any race, sex or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact. . ."

This definition was adopted as a practical "rule of thumb" and not a legal definition.

[1] Prepared by S. B. Sells

In the selection procedures for hiring applicants for the air traffic control occupation, adverse impact can occur at two points in the selection process. The first point involves basic eligibility for employment consideration, which depends on achievement of a passing score on the test battery. The second point involves the Earned Rating score (which includes extra awarded credit for veterans preference and aviation-related experience or knowledge); the Earned Rating determines the competitive ranking of applicants who pass the written test.

The information base that was employed to assess adverse impact consisted of data on the group of 5,976 applicants who completed the CSC battery and the proposed experimental test battery in 1978 (See Chapter 20). No information on race-ethnic group or sex has been obtained by OPM for other groups of applicants who completed both test batteries.

For this group of 1978 ATC applicants, the relative selection (pass) rates by race-ethnic group and sex, on the CSC battery are provided in Table 1. Comparable data for the experimental battery (WT$_2$) are presented in Table 2. Both tables exclude extra credit for veterans preference or aviation knowledge, since the purpose was to evaluate the relative rates of basic eligibility status of the applicants. These selection rates identify the proportion of each sex and race group who passed the selection test battery with a score of 70 or more. Based on these rates, the 80 percent "rule of thumb" was applied to each group in relation to the majority group selected, to derive the "adverse impact ratio" that would occur at this point in the selection process. When this ratio is less than 80 percent, the Uniform Guidelines state that this will generally be regarded as evidence of adverse impact. The Guidelines also point out that where the ratio is less than 80 percent, it may not constitute adverse impact, provided: (1) the differences are based on small numbers and are not statistically significant; or (2) special recruiting or other programs caused the pool of minority or female candidates to be atypical of the normal group of applicants. Since these 1978 applicants included a "walk-in" group, resulting from special recruiting efforts for minorities and women, the adverse impact analysis was performed for the total group (including 664 walk-in applicants), as well as the scheduled group, which would be representative of the normal pool of applicants.

Based on the data in Tables 1 and 2, the inclusion of the walk-in applicant group had relatively little effect on the ratios derived. Generally, there was a difference of only 1% between the Total Group and the Scheduled Group, except for Asians, for whom the group size was small.

The CSC battery had an adverse impact on both the Hispanic and black groups. In the light of past FAA experience in hiring minorities from competitive OPM registers, this result was not unexpected. However, Table 1 provides a quantitative measure of the degree of adverse impact for these two groups. Table 2 shows that the experimental battery had some adverse

504.

Table 1

Adverse Impact Analysis
CSC Test Battery

Group	Total Group			Total Group		
	Total Applicants	Pct. Who Passed	Adverse Impact Ratio	Scheduled Applicants	Pct. Who Passed	Adverse Impact Ratio
Men	4191	53	--	3835	54	--
Women	1785	45	85	1473	46	85
	5976			5308		
White	4067	64	--	3775	63	--
Hispanic	339	38	59	271	38	60
Black	1407	19	30	1116	18	29
Asian	57	61	95	50	56	89
Amer. Indian	61	51	80	58	50	79
	5931			5300		

505.

Table 2

Adverse Impact Analysis
Experimental Test Battery (WT$_2$)

Group	Total Group			Total Group		
	Total Applicants	Pct. Who Passed	Adverse Impact Ratio	Scheduled Applicants	Pct. Who Passed	Adverse Impact Ratio
Men	4191	56	--	3835	57	--
Women	1785	43	77	1473	45	78
	5931			5308		
White	4067	66	--	3775	67	--
Hispanic	339	41	62	271	41	61
Black	1407	14	21	1116	14	21
Asian	57	61	92	50	56	84
Amer. Indian	61	53	80	58	53	79
	5931			5300		

impact on women applicants, in addition to the Hispanic and black groups, and that the degree of adverse impact on the black group was somewhat greater than for the CSC battery. For American Indians, the addition of one or two more applicants to the pass group would have resulted in a selection ratio of 80% or more on either battery. Consequently, the evidence for adverse impact based on this sample is marginal, at best.

The analysis was extended for those who passed both batteries to determine whether there was adverse impact in each of three score range groups. Condidates who passed each test battery were ranked on competitive OPM registers based on their total scores and those with the highest scores were selected first. The Earned Rating, which includes extra credit for OKT score and veterans preference, was used for the analysis. The results are shown in Table 3 for the CSC battery, and Table 4, for the experimental battery (WT_2).

While there was no evidence of adverse impact in the score ranges of 70-89 for women, both the CSC battery and the experimental battery had adverse impact in the score range of 90 or more. As discussed in Chapter 20, a significant factor for women has been the extra credit granted for veterans preference and, to a lesser extent, for aviation-related knowledge, as measured by the Occupational Knowledge Test (OKT). It is also evident that the experimental battery resulted in a greater adverse impact on women in the 90 and above score range than did the CSC battery. This is primarily a result of the fact that raw test scores for all applicants were lower on the experimental battery, compared to the CSC battery. Relative to men, women scored lower on the weighted experimental battery. Indeed, women comprised 27% of the group with raw scores of 90 or more on the CSC battery, but comprised only 12% of this group on the weighted experimental battery (See Chapter 20).

With respect to race-ethnic groups on the two test batteries, Tables 3 and 4 show a "?" in some of the score ranges. In these cases, the number of applicants was small and the addition or subtraction of one or two applicants in the score range would have resulted in either meeting or not meeting the "80% rule of thumb." Consequently, evidence of adverse impact was not clear based on the sample analyzed.

It is clear that there was adverse impact on the black group . both batteries in the score range of 90 or higher. In the score range of 80-89 on the weighted experimental battery, and possibly on the CSC battery, there was also adverse impact on blacks. Again, this was primarily a result of the fact that as a group, blacks who passed the test scored lower than the majority group. In the case of Hispanics, there was possible adverse impact in the score range of 90 or higher on both batteries, but not in the range of 80-89 for the CSC battery.

In terms of competitive employment opportunities, based on the CSC battery, between 80% and 85% of the appointments were generally made from among applicants with Earned Ratings of 90 or higher. Consequently,

507.

Table 3

Adverse Impact Analysis
CSC Battery Earned Ratings

	Total 1978 Applicants Who Passed the CSC Battery	OPM Earned Rating 70-79			OPM Earned Rating 80-89			OPM Earned Rating 90+		
		N	Pct. Who Passed	Adverse Impact	N	Pct. Who Passed	Adverse Impact	N	Pct. Who Passed	Adverse Impact
Men	2236	718	32	--	796	36	--	722	32	--
Women	799	358	45	No	290	36	No	151	19	Yes
	3035	1076			1086			873		
White	2556	831	33	--	934	36	--	791	31	--
Hispanic	128	50	39	No	47	37	No	31	24	?
Black	264	161	61	No	77	29	?	26	10	Yes
Asian	35	12	34	No	8	23	Yes	15	43	No
Amer. Indian	31	10	32	No	15	48	No	6	20	?
	3014	1064			1081			869		

508.

Table 4

Adverse Impact Analysis
Experimental Battery (WT$_2$) Earned Ratings

	Total 1978 ATC Applicants Who Passed the CSC Battery	OPM Earned Rating 70-79			OPM Earned Rating 80-89			OPM Earned Rating 90+		
		N	Pct. Who Passed	Adverse Impact	N	Pct. Who Passed	Adverse Impact	N	Pct. Who Passed	Adverse Impact
Men	2348	951	40	--	905	39	--	492	21	--
Women	768	475	61	No	257	34	No	36	5	Yes
	3116	1426			1162			528		
White	2698	1171	43	--	1047	39	--	480	18	--
Hispanic	136	76	56	No	39	29	Yes	21	15	?
Black	198	134	68	No	50	25	Yes	14	7	Yes
Asian	35	16	46	No	13	37	No	6	17	No
Amer. Indian	32	19	59	No	9	28	?	4	13	?
	3099	1416			1158			525		

509.

employment prospects of applicants with scores in the 80-89 range on the experimental battery would have been "fair" to "good."

In the score range of 80 or higher on the weighted experimental battery, there was adverse impact for women and blacks. For the Hispanics and American Indians, the numbers were small and the addition of one or two more applicants within the score range would have met the "80%" criterion. These results are shown in Table 5. This analysis shows that the weighted experimental battery had adverse impact on women and on some minority groups, with respect to both initial eligibility (passing the test) and Earning Ratings. In this situation, the Uniform Guidelines require evidence of the validity of tests, the basis for establishing the pass cutoff score, and the use of scores for rank ordering of candidates.

Test Validity, Pass Score, and Candidate Ranking

The validity of the experimental battery, as well as the individual component tests has been demonstrated in relation to a number of ATC performance criterion measures. Briefly, these include:

The 1977 EPA Study (Mies, Colmen, and Domenech, 1977; See Chapter 18, Part II). For a sample of apporoximately 2,100 full-performance, developmental and newly hired ATC specialists, the MCAT, Directional Headings Test, OKT, and the Prior Experience Questionnaire (PEQ) provided significant correlations with an aggregate ATC success criterion for each of four ATC options (FSS, VFR, IFR, and ARTCC) and for all options combined. In addition, mean test scores for each of the four ATC options were different at statistically significant levels of confidence; the FSS option average score was lowest and the ARTCC score, highest.

The 1976-1978 ATC Study (Boone, 1979a; See Chapter 18, Part III). For a sample of 1,827 ATC trainees, experimental forms of MCAT, together with Abstract Reasoning (CSC 157) and Arithmetic Reasoning (CSC 24), provided a multiple correlation (corrected for range restriction) of .54 with the ATC laboratory problem average score. Cross validation, using the weighted test scores derived from the multiple regression, provided a multiple correlation (R) scores on the OKT were combined with the experimental test battery (MCAT, OPM 157 and OPM 24), an estimated multiple correlation value of .60 was obtained; the increase was significant at the .01 level of confidence.

The 1978 Validation Study of the Experimental Battery (See Chapter 21). For a new sample of 585 ATC trainees, parallel forms of MCAT together with CSC tests 157 and 24, provided an uncorrected multiple correlation (R) of .42 for the total sample. The R for sex and race-ethnic groups ranged from .37 to .69. The criterion used was the pass-fail status of the trainee at the completion of ATC Laboratory training. Validity coefficients used in the multiple regression were not corrected for range restriction. When

510.

Table 5

Adverse Impact Analysis
Experimental Battery (WT_2) for Earned Ratings of 80 or higher.

| | Total Applicants Who Passed the Exper. Battery | Earned Rating 80 or More | | Adverse Impact |
		N	Pct. Who Passed	
Men	2348	1397	60	--
Women	768	293	38	Yes
	3116	1690		
White	2698	1527	57	--
Hispanic	136	60	44	?
Black	198	64	32	Yes
Asian	35	19	54	No
Amer. Indian	32	13	41	?
	3099	1683		

511.

test scores on parallel forms of OKT were combined with the weighted test battery, the multiple correlation value for the total sample was .48, and R for sex and race-ethnic groups ranged from .44 to .76. The increase in the variance (R^2) accounted for by the weighted test battery with OKT included was significant at the .01 level of confidence for the total sample and for each race-ethnic group. For women, the increase was significant at the .05 level.

The utility of the experimental test battery, compared with that of the CSC battery, was examined with respect to appointment eligibility (pass scores) and ranking of successful candidates based on total scores including credit for veterans preference and ATC-related knowledge. For this purpose, the information available for the 1978 ATC trainee sample was related to the 1978 ATC applicant group. A common test battery, consisting of parallel forms of MCAT, CSC 157, and CSC 24, was administered to both groups. For the OKT, a 60-item test (101-C) was administered to the 1978 ATC applicant group and parallel forms of OKT 102 (80 item) administered to most of the 1978 ATC trainee sample. The common test battery was weighted:

> 4 x MCAT Total Right Scores
> 2 x CSC 157 Total Right Scores
> 1 x CSC 24 Total Right Scores

This weighting differed somewhat from the weights derived for the test battery that included two parallel forms of MCAT for the 1978 ATC trainee group. However, the correlation coefficient between the scores based on the two sets of weights was .93.

In the analysis of the experimental test battery, the passing score was set at the mean raw score for the 1978 ATC applicant group (222.27). This was then equated to a transmuted passing score of 70. With this base, equivalent scores on the same weighted test battery were computed for 585 of the 1978 trainees. A total of 48 trainees (8.2%) failed to achieve a score of 70. Of these 48 trainees, 79% (38) failed the ATC Laboratory training. Some 39 of the 48 trainees had scores in the 65-69 range and 9 scored below 65. The score range of 70-74 on the experimental battery included 102 of the 585 trainees (17%). Of this group, 51% (52) failed in training. Consequently, use of the mean test score resulted in a 28% reduction in the fail rates from those who scored in the 70-74 range to those who scored below 70 on the battery. The difference between a fail rate of 51% in the 70-74 score range and a rate of 79% for scores below 70 clearly supports the "cutoff" score developed in the analysis.

The use of scores on the experimental battery, with and without extra credit for veterans preference and ATC-related knowledge, was also examined as a basis for ranking successful ATC candidates. Since extra credit is a selection consideration only for those candidates who are appointed from

512.

competitive OPM registers, the analysis included only those 1978 ATC trainees who were hired competitively (non-predevelopmental) and who scored 70 or above on the experimental battery. On this basis, 509 of 585 trainees were included in the analysis. For the experimental battery without extra credit, the fail rate in each score range decreased progressively as the test scores increased. In the 95-100 range there were no failures (only two trainees were in this range). In the 70-74 range, 47.8% failed. The largest decrease in the fail rate was between the 75-79 range (41.9%) and the 80-85 range (29.3%), a difference of 12.6% in the fail rate. Table 6 combines the score ranges and provides the overall failure rates for this sample of 509 trainees. These data show clearly that ranking of ATC candidates according to their test scores and selecting those with the higher scores first, results in the hiring of applicants with the highest probability of successful completion of the ATC training program.

Ranking of candidates based on total Earned Rating scores (which include test scores plus extra credit for veterans preference and ATC-related knowledge based on OKT scores) was also examined. For the sample of 509 ATC trainees who were hired competitively, the equivalent of the OPM Earned Rating for the ATC test could be computed for 486 trainees. Table 7 shows the fail rates in training by Earned Rating score ranges and compares these with the rates based on the experimental battery score ranges without extra credit. It is evident from Table 7 that the ranking of candidates for selection consideration on the basis of the Earned Rating (which includes extra credit) also results in the selection of those applicants first with the highest probability of successful completion of training. It is also evident that selection based on Earned Rating results in a higher fail rate in each score range than if candidates had been hired solely on the basis of their test battery scores, without extra credit.

In deriving the Earned Rating score for this sample of 486 ATC trainees, extra credit for veterans preference was included since it is granted by law. Extra credit based on OKT test scores was included based on the positive and significant correlations with ATC performance criteria in the various research studies reported in earlier chapters. Consequently, it was hypothesized that extra credit for being a veteran, which has not been examined for validity with the criterion measures used in the research, does not predict (or negatively predicts) success in the ATC Laboratory training. In order to examine this, scores for the 509 trainees were computed, which excluded veterans credit and included only the weighted ATC test battery score and extra credit for OKT scores. Table 8 provides the score distributions for the 509 trainees, before and after the application of extra credit based on OKT.

It is significant that the addition of OKT points resulted in moving 110 trainees to a score range of 90 or higher -- an increase of about 22%. It also resulted in reducing the number of trainees in the lowest scoring

513.

Table 6

1978 ATC Trainees
ATC Test Scores
(Without Extra Credit)

Score Range	N	Number Failed	Fail Rate
90+	29	5	17.2%
80-89	254	66	26.0%
70-79	226	100	44.2%
Total	509	171	33.6%

Table 7

Fail rates, by score range on the Experimental
Battery, for 1978 trainees, with and without
extra credit.

Score Range	Experimental Battery (Without Extra Credit)			Experimental Battery (With Extra Credit)		
	N	Number Failed	Fail Rate	N	Number Failed	Fail Rate
90+	29	5	17.2%	197	39	19.8%
80-89	254	66	26.0%	229	86	37.6%
70-79	226	100	44.2%	60	38	63.3%
Total	509	171	33.6%	486	163	33.5%

Table 8

Distribution of scores on the Experimental Battery
without extra credit and with extra credit for
OKT only.

Score Range	Distribution Without OKT Credit		Distribution with OKT Credit	
	N	%	N	%
90+	29	5.7%	139	27.3%
80-89	254	49.9%	245	48.1%
70-79	226	44.4%	125	24.6%
Total	509	100.0%	509	100.0%

group (70-79) by 91 -- from 226 to 125 -- a reduction of about 20%. Based on the validity data presented for the OKT (Chapter 16), it would be expected that those trainees who moved to the higher score ranges as a result of OKT extra credit would have lower fail rates and that those who remained in the lowest score range would have higher fail rates. The comparison of the fail rates in the three score ranges for the 509 ATC trainees before and after applying OKT extra credit points is presented in Table 9.

Since the fail rate for the total group of 509 trainees was a constant (33.6%), the rates could vary only between the different score ranges, as shown in Table 9. In order to assess the effect of OKT extra credit on the total fail rate, one can assume a selection made from among the group of 509 trainees. For example, if one were to hire 300 trainees from the 509 in rank order (highest score range first) with an estimated selection ratio of 80 in each score range, the fail rate for the 300 hired before application of OKT extra credit would be 30%. The fail rate for the 300 selected after application of OKT extra credit would be 25%. If only 100 of the 509 were selected, the fail rate for the group selected without OKT extra credit would be 24% in contrast to 14% for the group with OKT credit. These data support the conclusion that ranking of applicants for selection consideration based on test scores and OKT credit will result in hiring applicants with considerably higher probability of successful completion of ATC training.

Table 10 shows the fail rates for the 509 trainees based on (1) the experimental battery alone; (2) the experimental battery plus OKT credit; and (3) the experimental battery plus credit for OKT and veterans preference points. The use of OKT credit resulted in moving 110 trainees into the 90 and above score range, compared to the experimental battery scores alone. The number of fails increased by 14, for a rate of 13.7% in this group. When veterans preference credit was also applied, for 486 of the 509 trainees, an additional 58 trainees moved to the 90 and above score range and the number of fails increased by 20, giving a fail rate of 34% in this group of 39 students. Thus from the data presented in Table 10, the granting of extra credit for veterans preference increased the fail rates in the higher score ranges and therefore had a negative relationship to potential success in ATC training.

Test Fairness

When empirical evidence of validity for a selection procedure, such as presented for the experimental battery, has an adverse impact on a race, sex, or ethnic group, the Uniform Guidelines require examination of test fairness, where technically possible.

In establishing the Uniform Guidelines on Employee Selection, it was recognized there is serious debate on the question of test fairness. There are several competing definitions of test or selection fairness, each of which incorporates a different set of social values. Hunter and

517.

Table 9

Fail rates by score range on the Experimental
Battery, with and without OKT credit.

Score Range	Without OKT Credit			With OKT Credit		
	N	Number Failed	Fail Rate	N	Number Failed	Fail Rate
90+	29	5	17.2%	139	19	13.7%
80-89	254	66	26.0%	245	78	31.8%
70-79	226	100	44.2%	125	74	59.2%
Total	509	171	33.6%	509	171	33.6%

Table 10

Fail rates by experimental battery score ranges, (1) with no extra credit, (2) OKT extra credit only, and (3) with OKT extra credit and veterans preference credit. N = 509 - 1978 trainees.

Score Range	Exper. Battery Scores with No Extra Credit				Exper. Battery Scores plus OKT Extra Credit Only				Exper. Battery Scores plus OKT and Vet. Pref. Credit			
	N	Percent	No. Fail	Fail Rate (Pct.)	N	Percent	No. Fail	Fail Rate (Pct.)	N	Percent	No. Fail	Fail Rate (Pct.)
90+	29	6	5	17.2	139	27	19	13.7	197	41	39	19.8
80-89	254	50	66	26.0	245	48	78	31.8	229	47	86	37.6
70-79	226	44	100	44.2	125	25	74	59.2	60	12	38	63.3
Total	509	100	171	33.6	509	100	171	33.6	486	100	163	33.5

Schmidt (1976) provided an analysis of three incompatible ethical positions in regard to fair and unbiased use of tests, together with differing statistical definitions of "test fairness" and their relationship to specific ethical positions. These three ethical definitions are characterized as Unqualified Individualism, Qualified Individualism, and the Quota Ethic. A report of the General Accounting Office (1979) evaluated Federal selection tests and examinations, and states with respect to these three definitions:

> "The published literature on test validity indicates that most tests are either fair to minority groups or slightly biased in their favor by the second definition, (Qualified Individualism) which according to OPM, is the only concept of fairness consistent with merit system principles. THe literature also indicates that by the first definition, (Unqualified Individualism) tests are slightly biased against minority groups, and if one subscribes to the last definition (Quota Ethic) tests have always been biased against minorities."

The Qualified Individualism definition of fairness adopted by the Office of Personnel Management (OPM) holds that tests are biased when those with equal chances of success on the job have unequal chances of being selected for the job. This definition relies solely on valid measures of aptitude, achievement, and experience; it maximizes productivity, and provides equal opportunity consistent with merit principles. The statistical model upon which the Uniform Guidelines appear to be based, was stated by Cleary in Hunter and Schmidt (1976). Basically, this model requires that the regression line which predicts the criterion (e.g. performance) from predictive scores to be the same for all cultural groups and, in the absence of this condition, that separate regression lines should be used as a basis for selection decisions. Statistical tests developed by Gulliksen and Wilkes (1970) provide a procedures which meets the Cleary model. Basically, three independent and sequential tests are involved:

(1) population variances (standard error of estimate) are equal (non-significant differences, then,

(2) slopes of the population regression lines are equal (non-significant differences), then,

(3) intercepts of the population regression lines are equal (non-significant differences).

If each of the three statistical tests is met, the regression lines are considered to be the same and the fairness definition of Qualified Individualism is satisfied, in that predictor scores used for selection predict the criterion equally for the various groups considered. On the other hand, if a statistically significant difference is found at any point in the sequential analysis, the test is considered to indicate that the groups were treated unequally.

In order to examine the experimental battery in accordance with the Qualified Individualism model of fairness, test scores and criterion data

obtained for the sample of 953 ATC trainees attending the FAA Academy during June 1978 through December 1978 were used in the analysis. The number of Hispanics (23 total -- 7 with CSC test data) was too small for analysis. There were no American Indians in this sample and there was no evidence of adverse impact on selection of Asians with the experimental battery. Consequently, the analysis for test fairness was completed with respect to men and women and for white and black trainees.

Sample sizes varied depending on the availability of data on the two CSC tests (CSC 157 and 24) included in the experimental battery. Test and criterion data were available for ATC trainees, as shown in Table 11. In this analysis, the tests were weighted in accordance with the values derived for the 1978 ATC trainees. These weights were:

$$\left. \begin{array}{ll} 2 \text{ x MCAT 1} & \text{(Total Right Scores)} \\ 2 \text{ x MCAT 2} & \text{(Total Right Scores)} \end{array} \right) = \text{MCAT Total x 2}$$

1 x CSC 157   (Total Right Scores)
1 x CSC 24   (Total Right Scores)

The test for significance for all analyses of fairness was the .05 level of confidence.

Since almost all of the 1978 ATC trainee sample had test scores available on the two parallel forms of MCAT, an analysis of fairness of the MCAT by itself was completed. Table 12 shows the results for the first form of the test administered (MCAT 1) and the total score for both forms (MCAT 1 plus MCAT 2) by race-ethnic group and sex. The test means and standard deviations were based on raw test scores. Since MCAT was the only test used in this analysis, the raw scores were not weighted. The criterion value was based on a mean of 0 with a standard deviation of 1 and reflects, in effect, the percent of trainees who passed the Laboratory training. The validity coefficient (r) is the correlation of the test with the pass-fail criterion. With reference to the size of samples used in this analysis, the 137 women comprised 14.8% of the combined sample of men and women (927) compared to 15.7% of the 7,894 ATC trainees hired from January 1976 through October 1980, who were women. The 79 black trainees represented 8.7% of the combined black and white sample (907), compared to 8.3% (of blacks) among the 7488 black and white ATC trainees hired during this same period.

For men and women, the test of fairness showed no significant differences (NS). For the white and black groups, the difference in the intercept of the regression line, was statistically significant (SIG) for the first MCAT test administered, but differences in the standard error and the slope of the regression line were not significant (NS). When the total scores for both MCAT forms administered were analyzed, there were no significant differences between the two groups. This analysis indicates that the total MCAT scores predicted the ATC laboratory pass-fail criterion equally for men and women and for the black and white groups.

Table 11

Summary of test data available for analysis of fairness of the experimental battery by the Qualified Individualism model.

	Men (N=800) Sample N	Women (N=141) Sample N	White (N=839) Sample N	Black (N=81) Sample N
Test				
MCAT 1	790	137	828	79
MCAT TOTAL	790	137	828	79
CSC 157	509	66	539	38
CSC 24	509	66	539	38

522.

Table 12

Analysis of test fairness of the MCAT
for Men vs Women and Whites vs Blacks.

Men

Test	N	Test Mean	SD	Criterion Mean	SD	r
MCAT 1	790	37.38	6.86	.63	.48	.34
MCAT TOTAL	790	80.30	11.25	.63	.48	.40

Women

Test	N	Test Mean	SD	Criterion Mean	SD	r	Test of Significance Std. Error	Slope	Inter-cept
MCAT 1	137	34.12	7.82	.58	.50	.37	NS	NS	NS
MCAT TOTAL	137	74.78	13.12	.58	.50	.41	NS	NS	NS

White

Test	N	Test Mean	SD	Criterion Mean	SD	r
MCAT 1	828	37.57	6.70	.65	.48	.29
MCAT TOTAL	828	80.73	10.75	.65	.48	.35

Black

Test	N	Test Mean	SD	Criterion Mean	SD	r	Test of Significance Std. Error	Slope	Inter-cept
MCAT 1	79	29.75	7.87	.35	.48	.56	NS	NS	SIG
MCAT TOTAL	79	66.36	13.22	.35	.48	.57	NS	NS	NS

523.

The next step was to develop different test battery combinations with the MCAT test scores and CSC 157, CSC 24, and OKT test scores and to analyze these various test batteries for significant differences between men and women and the white and black groups. The intercorrelations between the various test combinations used and their correlation with the pass-fail criterion are shown for each group in Table 13. Since a number of trainees did not have test scores on CSC 157 and CSC 24, the sample sizes were reduced as shown in Table 13. The correlations shown are based on weighted test scores, where appropriate, using raw test battery scores (RS) as well as transmuted scores (TS). The correlations which combined a test battery with OKT used both the raw OKT score (RS) and OKT extra point scores (PTS), derived from the raw OKT score (i.e., 0, 3, 5, 10, 15 points).

The descriptive statistics for the samples used in the fairness analysis are also provided in comparison to the total sample for each group as was given in Table 6, Chapter 21.

The first test battery analyzed for fairness consisted of the MCAT Total and CSC 175 raw scores. In this analysis, MCAT Total was weighted by 2 and CSC 157 weighted by 1 in accordance with the weights previously derived. Then the OKT raw scores (weighted 1) were combined with the MCAT and CSC 157 raw scores and this combination was evaluated separately. The results for each sex and race-ethnic group are shown in Table 14. It should be noted that at this point, a passing score for the combined MCAT and CSC 157 test scores had not been established. Consequently, the addition of the OKT raw score was independent of whether or not the ATC trainee passed the MCAT/CSC 157 test battery and also independent of extra credit points (0, 3, 5, 10, 15) which could be derived from the raw OKT scores.

The combination of MCAT and CSC 157 raw test scores predicted the ATC laboratory pass-fail criteria equally for men and women and for black and white groups. The same results were obtained for the battery consisting of MCAT, CSC 157, and OKT raw test scores.

Use of raw scores in analyzing the test battery composed of MCAT, CSC 157, and OKT did not take into account the proposed use of OKT in the selection process. Since OKT is a job-specific test, it was not intended for use in the initial pass-fail determination of appointment eligibility for ATC applicants. Its use was intended to be limited to determining whether those applicants who passed the test battery (in this case MCAT and CSC 157) would also be given extra credit for air traffic control-related knowledge as measured by OKT. Further, the amount of extra credit (0, 3, 5, 10, 15 points) was to be based on the raw OKT score with no credit for raw scores below 65 and maximum credit (15 points) for raw OKT scores of 80 or above.

In order to evaluate the fairness of a test battery in combination with OKT in relation to its intended use, it was necessary to (1) establish

524.

Table 13

Sample and population means, standard deviations, and intercorrelations, for men and women (Part A), and for black and white trainees (Part B), for the MCAT (1, 2, and total), SC 157, CSC 24, and OKT. Intercorrelations of (1) MCAT Total, (2) CSC 157, (3) CSC 24, (4) OKT (RS - raw scores), (5) MCAT Total + 157 (RS), (6) MCAT Total + 157 (TS - transmuted scores), (7) MCAT Total + 157 + OKT (RS), (8) MCAT Total + 157 (TS) + OKT (PTS - extra points), (9) MCAT Total + 157 + 24 (RS), (10) MCAT Total + 157 + 24 + OKT (RS), (11) MCAT Total + 157 + 24 (TS), (12) MCAT Total + 157 + 24 (TS) + OKT (PTS), (13) Criterion (Lab Pass-Fail).

Part A. Men vs Women

| | Men | | | | | | Women | | | | | |
| | Sample | | | Population | | | Sample | | | Population | | |
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
MCAT 1	509	37.6	6.6	800	37.4	6.9	66	33.4	8.1	141	34.1	7.8
MCAT 2	509	43.3	5.7	800	42.9	5.8	66	40.7	6.7	141	40.7	6.4
MCAT TOT.	509	80.9	11.0	800	--	--	66	74.1	13.4	141	--	--
CSC 157	509	38.8	6.2	515	38.8	6.2	66	39.9	6.0	67	39.8	6.0
CSC 24	509	46.7	6.5	515	46.6	6.5	66	47.3	7.1	67	47.2	7.1
OKT (RS)	509	64.1	15.8	800	64.9	15.5	66	56.5	14.4	141	57.7	15.1

Table 13 (Continued)

Men
(N=509)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	MCAT TOT.	CSC 157	CSC 24	OKT (RS)	MCAT TOT. 157 (RS)	MCAT TOT. 157 (TS)	MCAT TOT. 157 OKT (RS)	(TS) MCAT TOT. 157 OKT (PTS)	MCAT TOT. 157 24 (RS)	MCAT TOT. 157 24 OKT (RS)	MCAT TOT. 157 24 (TS)	(TS) MCAT TOT. 157 24 OKT (PTS)	Pass-Fail
(1) MCAT TOT.	--	.24	.33	.03	.97	.97	.83	.82	.95	.86	.95	.82	.38
(2) CSC 157		--	.09	-.16	.48	.48	.32	.32	.45	.32	.45	.30	.16
(3) CSC 24			--	-.23	.32	.32	.14	.15	.53	.35	.53	.33	.12
(4) OKT(RS)				--	-.02	-.02	.54	.44	-.07	.46	-.07	.41	.21
(5) MCAT TOT. + 157(RS)					--	1.00	.83	.83	.97	.86	.97	.82	.38
(6) MCAT TOT. + 157(TS)						--	.83	.83	.97	.86	.97	.82	.38
(7) MCAT TOT. + 157 + OKT(RS)							--	.94	.78	.98	.78	.92	.44
(8) MCAT TOT. + 157(TS) + OKT(PTS)								--	.78	.92	.78	.98	.44
(9) MCAT TOT. + 157 + 24(RS)									--	.85	1.00	.82	.37
(10) MCAT TOT. + 157 + 24 + OKT(RS)										--	.85	.94	.44
(11) MCAT TOT. + 157 + 24(TS)											--	.82	.37
(12) MCAT TOT. + 157 + 24(TS) + OKT(PTS)												--	.44
(13) CRITERION (Lab Pass-Fail)													--

Table 13 (Continued)

Women

(N=66)

	(1) MCAT TOT.	(2) CSC 157	(3) CSC 24	(4) OKT (RS)	(5) MCAT TOT. 157 (RS)	(6) MCAT TOT. 157 (TS)	(7) MCAT TOT. 157 OKT (RS)	(8) (TS) MCAT TOT. 157 OKT (PTS)	(9) MCAT TOT. 157 24 (RS)	(10) MCAT TOT. 157 24 OKT (RS)	(11) MCAT TOT. 157 24 (TS)	(12) (TS) MCAT TOT. 157 24 OKT (PTS)	(13) Pass-Fail
(1) MCAT TOT.	--	.23	.04	.17	.98	.98	.89	.92	.95	.89	.95	.90	.33
(2) CSC 157		--	.12	.06	.42	.42	.38	.34	.43	.40	.43	.35	.40
(3) CSC 24			--	-.25	.06	.06	-.06	.01	.29	.15	.29	.20	.08
(4) OKT(RS)				--	.17	.17	.56	.41	.11	.51	.11	.37	.31
(5) MCAT TOT. + 157(RS)					--	1.00	.91	.93	.97	.91	.97	.91	.39
(6) MCAT TOT. + 157(TS)						--	.91	.93	.97	.91	.97	.91	.39
(7) MCAT TOT. + 157 + OKT(RS)							--	.95	.86	.98	.86	.92	.46
(8) MCAT TOT. + 157(TS) + OKT(PTS)								--	.89	.95	.89	.98	.41
(9) MCAT TOT. + 157 + 24(RS)									--	.91	1.00	.92	.39
(10) MCAT TOT. + 157 + 24 + OKT(RS)										--	.91	.95	.47
(11) MCAT TOT. + 157 + 24(TS)											--	.92	.39
(12) MCAT TOT. + 157 + 24(TS) + OKT(PTS)												--	.42
(13) CRITERION (Lab Pass-Fail)													--

527.

Table 13 (Continued)

Part B. White vs Black

| | White | | | | | | Black | | | | | |
| | Sample | | | Population | | | Sample | | | Population | | |
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
MCAT 1	539	37.7	6.6	839	37.6	6.7	38	29.7	7.8	81	29.8	7.9
MCAT 2	539	43.6	5.5	839	43.2	5.5	38	36.1	7.5	81	36.6	6.8
MCAT TOT.	539	81.2	10.6	839	--	--	38	65.8	14.0	81	--	--
CSC 157	539	39.1	6.2	545	39.1	6.2	38	35.4	6.1	39	35.3	6.1
CSC 24	539	66.8	6.6	545	46.9	6.5	38	44.7	6.6	39	44.2	7.1
OKT(RS)	539	63.1	15.9	839	63.9	15.8	38	65.5	13.6	81	66.2	13.1

r

Table 13 (Continued)

White

(N=539)

	(1) MCAT TOT.	(2) CSC 157	(3) CSC 24	(4) OKT (RS)	(5) MCAT TOT. 157 (RS)	(6) MCAT TOT. 157 (TS)	(7) MCAT TOT. 157 OKT (RS)	(8) (TS) MCAT TOT. 157 OKT (PTS)	(9) MCAT TOT. 157 24 (RS)	(10) MCAT TOT. 157 24 OKT (RS)	(11) MCAT TOT. 157 24 (TS)	(12) (TS) MCAT TOT. 157 24 OKT (PTS)	(13) Pass-Fail
(1) MCAT TOT.	--	.23	.28	.07	.97	.97	.83	.82	.94	.85	.94	.82	.33
(2) CSC 157		--	.10	-.14	.48	.48	.31	.32	.45	.32	.45	.31	.17
(3) CSC 24			--	-.24	.28	.28	.09	.10	.50	.31	.50	.30	.07
(4) OKT(RS)				--	.03	.03	.58	.68	-.03	.50	-.03	.44	.23
(5) MCAT TOT. + 157(RS)					--	1.00	.83	.82	.97	.86	.97	.82	.34
(6) MCAT TOT. + 157(TS)						--	.83	.82	.97	.86	.97	.82	.34
(7) MCAT TOT. + 157 + OKT(RS)							--	.94	.77	.96	.77	.92	.41
(8) MCAT TOT. + 157(TS) + OKT(PTS)								--	.77	.92	.77	.98	.40
(9) MCAT TOT. + 157 + 24(RS)									--	.85	1.00	.81	.33
(10) MCAT TOT. + 157 + 24 + OKT(RS)										--	.85	.94	.41
(11) MCAT TOT. + 157 + 24(TS)											--	.81	.33
(12) MCAT TOT. + 157 + 24(TS) + OKT(PTS)												--	.40
(13) CRITERION (Lab Pass-Fail)													--

529.

Table 13 (Continued)

Black

(N=38)

	(1) MCAT TOT.	(2) CSC 157	(3) CSC 24	(4) OKT (RS)	(5) MCAT TOT. 157 (RS)	(6) MCAT TOT. 157 (TS)	(7) MCAT TOT. 157 OKT (RS)	(8) (TS) MCAT TOT. 157 OKT (PTS)	(9) MCAT TOT. 157 24 (RS)	(10) MCAT TOT. 157 24 OKT (RS)	(11) MCAT TOT. 157 24 (TS)	(12) (TS) MCAT TOT. 157 24 OKT (PTS)	(13) Pass-Fail
(1) MCAT TOT.	--	-.28	.17	.27	.98	.98	.90	.92	.96	.91	.96	.91	.60
(2) CSC 157		--	-.11	-.11	-.06	-.06	-.10	-.18	-.09	-.12	-.09	-.19	.07
(3) CSC 24			--	-.09	.15	.15	.06	.09	.37	.27	.37	.25	.35
(4) OKT(RS)				--	.25	.25	.62	.49	.22	.58	.22	.47	.24
(5) MCAT TOT. + 157(RS)					--	1.00	.92	.92	.97	.91	.97	.90	.64
(6) MCAT TOT. + 157(TS)						--	.92	.92	.97	.91	.97	.90	.64
(7) MCAT TOT. + 157 + OKT(RS)							--	.95	.88	.96	.88	.93	.62
(8) MCAT TOT. + 157(TS) + OKT(PTS)								--	.89	.94	.90	.99	.61
(9) MCAT TOT. + 157 + 24(RS)									--	.92	1.00	.91	.68
(10) MCAT TOT. + 157 + 24 + OKT(RS)										--	.92	.95	.67
(11) MCAT TOT. + 157 + 24(TS)											--	.91	.68
(12) MCAT TOT. + 157 + 24(TS) + OKT(PTS)												--	.64
(13) CRITERION (Lab Pass-Fail)													--

530.

Table 14

Analyses of test fairness of MCAT Total + CSC 157 and MCAT Total + CSC 157 + OKT, for men vs women and whites vs blacks.

| | | Men | | | | | | Women | | | | | Test of Significance | | |
| | | Test | | | Criterion | | | Test | | | Criterion | | | | |
Test	N	Mean	SD	Mean	SD	r	N	Mean	SD	Mean	SD	r	Std. Error	Slope	Intercept
MCAT TOT.)															
CSC 157)) - (RS)	509	200.62	24.19	.629	.484	.38	66	188.02	28.80	.515	.504	.39	NS	NS	NS
MCAT TOT.)															
CSC 157) - (RS)															
OKT)	509	264.71	28.68	.629	.484	.44	66	244.47	34.33	.515	.504	.46	NS	NS	NS

| | | White | | | | | | Black | | | | | Test of Significance | | |
| | | Test | | | Criterion | | | Test | | | Criterion | | | | |
	N	Mean	SD	Mean	SD	r	N	Mean	SD	Mean	SD	r	Std. Error	Slope	Intercept
MCAT TOT.)															
CSC 157)) - (RS)	539	201.54	23.37	.640	.480	.34	38	166.93	26.88	.316	.471	.64	NS	NS	NS
MCAT TOT.)															
CSC 157) - (RS)															
OKT)	539	264.62	28.62	.640	.480	.41	38	232.41	33.03	.316	.471	.62	NS	NS	NS

(RS) - Raw Scores

531.

a "passing" score on the test battery, (2) transmute the raw test battery
score in relation to the passing score, and (3) add the appropriate extra
credit points based on OKT raw score, for those who passed the test bat-
tery.

In order to transmute the total raw score for the ATC trainee sample,
it was necessary to estimate the mean score which would have been obtained
by an ATC applicant group on this test battery (MCAT + CSC 157) and equate
this mean score to a passing score of 70. Fortunately, the tests used in
the analysis of the ATC trainee sample were identical to the tests taken
by the 5,931 ATC applicants in 1978 except for the second form of MCAT,
which was not administered to the ATC applicants. Consequently, given
the mean scores on each test, which were available for both the ATC train-
ees and the ATC applicants, it was possible to estimate a mean for the
ATC applicants on the second MCAT form. By combining this estimated mean
for MCAT 2 with the actual mean scores obtained by the ATC applicants on
the other tests and weighting them, an estimated mean score of 161.2 for
the 1978 ATC applicant group was derived for a test battery composed of
MCAT Total and CSC 157, as shown below:

	1978 ATC Applicants				1978 ATC Trainees		
	MEAN	x	WT.	= TOTAL	MEAN	x	WT.	= TOTAL
MCAT 1	30.6	·	2	61.2	36.9		2	73.8
*(MCAT2	35.6		2	71.2)	42.6		2	85.2
CSC 157	28.8		1	28.8	38.8		1	38.8
Est. Mean Raw Score				161.2				197.8

(1) ATC Applicant Mean – MCAT 1 $\dfrac{30.6}{36.9}$ = .829
     ATC Trainee Mean – MCAT 1

(2) ATC Trainee Mean – MCAT 2   42.6
     ATC Trainee Mean – MCAT 1   36.9
        Difference $\overline{5.7}$ (6.0)

*(3) Est. MCAT 2 Mean for Applicant Group = 35.6
      (6.0 x .829 = 5.0 + 30.6 = 35.6)

Since the correlation between the two differently weighted experimental
battery scores for the ATC trainee sample was .93 (Table 20, Chapter 21),
an estimated mean score for the ATC applicant group of 159.5 was also de-
rived, using the ratio of the mean scores for the 1978 applicant and train-
ee groups for the common weighted tests as shown below.

	MEAN

(1) Actual Weighted Applicant Raw Score Mean
     (N = 6000)   (4 x MCAT 1 + 2 x 157)   181.6

(2) Actual Weighted Trainee Raw Score Mean
     (N = 592)   (4 x MCAT 1 + 2 x 157)   225.2

(3) Actual Weighted Trainee Raw Score Mean
     (N = 592)   (2 x MCAT 1 + 2 x MCAT 2 + 1 x 157)   197.8

(4) $\dfrac{197.8}{225.2}$ x $\dfrac{x}{181.6}$ = 159.5 Est. MEAN for Applicant Group
                                 on ATC Test Weighted
                                 (2 x MCAT 1 + 2 x MCAT 2 + 1 x 157)

Given these two estimated mean scores for an applicant group on the ATC
test battery comprised of MCAT and CSC 157, a mean value of 160 was used
as a passing score and equated to a transmuted score of 70. The maximum
raw weighted score was equated to a transmuted score of 100. As each
person's raw weighted score was greater than 160, but less than the maxi-
mum possible raw weighted score, these scores were converted to fit the
new 30-point score range between 70 and 100 by the following general
transformation:

$$T_s = (\frac{(RS - 160) \times 30}{MAX\ RS = 160}) + 70$$

RS   = Total weighted raw score
MAX RS   = 255.0
MEAN RS = 160.0
70   = New Passing Score
30   = New Score Range (70 to 100)

An analysis of fairness was completed using the obtained transmuted test
battery scores for the ATC trainees. A separate analysis was completed
using the transmuted scores plus extra credit points (0, 3, 5, 10, 15)
derived from OKT raw scores for those ATC trainees who passed the test
battery with scores of 70 or more. The results are shown in Table 15.

The combination of transmuted test battery scores for MCAT and CSC
157 plus extra credit points for OKT scores replicates the operational
use of the selection tests. The analysis of fairness provided in Table
15 shows that this test battery including OKT, predicts the ATC laboratory
pass-fail criterion equally for men and women and for black and white
groups.

Finally, an examination was made of the fairness of the entire test
battery, composed of MCAT, CSC 157, and CSC 24 with and without extra
credit based on OKT scores. The results are displayed in Table 16. In

Table 15

Analyses of test fairness of a battery consisting of the MCAT Total, CSC 157, and OKT (raw score and extra points), for men vs women and whites vs blacks.

		Men			Criterion			Test of Significance					Women			Criterion			Test of Significance		
Test		Test Mean	SD	Mean	SD	r	Std. Error	Slope	Inter-cept			Test Mean	SD	Mean	SD	r	Std. Error	Slope	Inter-cept
	N										N								
MCAT TOT.) CSC 157) - (TS)	509	82.83	7.64	.629	.484	.38	NS	NS	NS		66	78.85	9.10	.515	.504	.39	NS	NS	NS
MCAT Tot.) CSC 157) - (TS) OKT) (PTS)	509	87.50	9.92	.629	.484	.44	NS	NS	NS		66	80.88	10.75	.515	.504	.41	NS	NS	NS

		White			Criterion			Test of Significance					Black			Criterion			Test of Significance		
Test		Test Mean	SD	Mean	SD	r	Std. Error	Slope	Inter-cept			Test Mean	SD	Mean	SD	r	Std. Error	Slope	Inter-cept
	N										N								
MCAT TOT.) CSC 157) - (TS)	539	83.12	7.38	.640	.480	.34	NS	NS	NS		38	72.19	8.49	.316	.471	.64	NS	NS	NS
MCAT TOT.) CSC 157) - (TS) OKT) (PTS)	539	87.48	9.68	.640	.480	.40	NS	NS	NS		38	76.14	11.90	.316	.471	.61	NS	NS	NS

(TS) - Transmuted Test Battery Score
(PTS) - OKT Points

534.

Table 16

Analyses of test fairness of a battery consisting of
MCAT Total, CSC 157, CSC 24, and OKT (raw score and
extra points), for men vs. women and whites vs blacks.

	Men						Women						Test of Significance		
		Test		Criterion				Test		Criterion					Inter-
Test	N	Mean	SD	Mean	SD	r	N	Mean	SD	Mean	SD	r	Std. Error	Slope	cept
MCAT TOT.)															
CSC 157) - (RS)	509	247.29	26.95	.629	.484	.37	66	235.27	30.08	.515	.504	.39	NS	NS	NS
CSC 24)															
MCAT TOT.)															
CSC 157) - (TS)	509	83.38	7.63	.629	.484	.37	66	79.98	8.51	.515	.504	.39	NS	NS	NS
CSC 24)															
MCAT TOT.))															
CSC 157) - (TS))															
CSC 24)) - (PTS)	- 509	88.04	9.72	.629	.484	.44	66	82.01	10.09	.515	.504	.42	NS	NS	NS
OKT)															

535.

Table 16 (Continued)

Test	White						Black						Test of Significance		
		Test		Criterion				Test		Criterion			Std.		Inter-
	N	Mean	SD	Mean	SD	r	N	Mean	SD	Mean	SD	r	Error	Slope	cept
MCAT TOT.)															
CSC 157) - (RS)	539	248.32	25.97	.640	.480	.33	38	211.62	28.61	.316	.471	.68	SIG	--	--
CSC 24)															
MCAT TOT.)															
CSC 157) - (TS)	539	83.67	7.35	.640	.480	.33	38	73.29	8.10	.316	.471	.68	SIG	--	--
CSC 24)															
MCAT TOT.))															
CSC 157) - (TS))															
CSC 24)) - 539	88.03	9.42	.640	.480	.40	38	77.16	11.45	.316	.471	.64	NS	NS	NS	
OKT) - (PTS)															

536.

order to add OKT extra credit points, it was again necessary to transmute
the total raw score for MCAT, CSC 157, and CSC 24 and to establish a pass-
ing raw score, equated to a transmuted score of 70, as shown below:

| | 1978 ATC Applicants | | | | 1978 ATC Trainees | | | |
TEST	MEAN	x	WT	=	TOTAL	MEAN	x	WT	=	TOTAL
MCAT 1	30.6		2		61.2	36.9		2		73.8
*(MCAT 2	35.6		2		71.2)	42.6		2		85.2
CSC 157	28.8		1		28.8	38.8		1		38.8
CSC 24	40.6		1		40.6	46.6		1		46.6
Est. Mean Raw Score					201.8					244.4

(*See earlier for derivation of estimated mean score for MCAT 2.)

Again, an estimated mean raw score for the applicant group was derived,
using a ratio method as follows:

(1) Actual Applicant Mean (N = 6000)
    (4 x MCAT 1 + 2 x 156 + 1 + 24)   = 222.3 (Table 28,
                                                                Chapter 20)

(2) Actual Mean ATC Trainee Group (N = 592)
    (4 x MCAT 1 + 2 x 157 + 1 x 24)   = 272.6 (Table 20,
                                                                Chapter 21)

(3) Actual Mean ATC Trainee Group (N = 592)
    (2 x MCAT 1 + 2 x MCAT 2 + 1 x 157 + 1 x 24)   = 245.6 (Table 20,
                                                                Chapter 21)

(4) $\frac{245.6}{272.6}$ $\frac{x}{222.3}$ = 200.3 Est. Mean for Applicant Group on ATC Test Weighted
    (2 x MCAT 1 + 2 x MCAT 2 + 1 x 157 + 1 x 24)

Given these two estimated mean scores for an applicant group (201.8 and 200.3)
on the experimental battery composed of MCAT, CSC 157, and CSC 24, a mean
value of 200 was used as a passing score and equated to a transmuted score
of 70. The total raw weighted scores for each ATC trainee on MCAT, CSC
157, and CSC 24 were then transmuted as follows:

$$T_s = \left(\frac{(RS - 200) \times 30}{MAX\ RS - 200} \right) + 70$$

R_s = Total weighted raw score

MAX RS = 306
MEAN RS = 200

Table 16 shows the results of the fairness analysis on this battery
(for both raw and transmuted scores) and for the experimental battery
transmuted scores in combination with extra credit points based on OKT,
for those trainees who scored 70 or more on the test battery.

As shown in Table 16, there were no significant differences on the experimental battery between men and women. However, as a result of the addition of CSC 24 test scores (Arithmetic Reasoning) to the scores of MCAT and CSC 157, the estimated population variances of errors of prediction between the black and white groups became statistically significant. The statistical tests developed by Gulliksen and Wilkes (1950) require that further tests be discontinued if a significant difference is found at any of the three steps in the analysis, since the regression lines used for prediction are, by definition, unequal.

The fact that the regression lines for the white and black groups were statistically unequal, by itself, does not identify the practical implications for the selection procedure. According to the Cleary model of test fairness, if tests are "biased" (unequal), an alternative solution is to use the separate regressions for selection decisions. However, adoption of this solution would have explicitly introduced race-ethnic group as a predictor in the selection process, and this is incompatible with the ethical position of Qualified Individualism as well as the law (Hunter and Schmidt, 1960).

In such circumstances, a single regression line derived by combining the groups typically has been used as the basis for selection procedures. The effect of this is to "overpredict" the criterion (e.g., job performance) for the group that scored lower on the test and had a lower mean on the criterion measure (Hunter and Schmidt, 1960). In the analysis of this test battery, use of a single regression line obtained by combining the white and black groups (the total sample) would "bias" the selection procedure in favor of the black group. This result conforms with the general findings in the published literature referred to by the General Accounting Office review of Federal selection tests (General Accounting Office, 1979).

Table 16 also shows that when extra credit points based on OKT were added to the transmuted scores for the experimental battery, there were no significant differences between men and women or the black and white groups. While this indicates that the combination of the test battery and OKT points predicts the ATC laboratory pass-fail criterion equally for all groups, the test battery itself would be used to establish initial appointment eligibility for applicants. The fact that addition of OKT points results in equal treatment for those who pass the test does not adequately address the unequal treatment resulting from the use of CSC 24 under the Uniform Guidelines. Given that CSC 24 results in unequal prediction of the pass-fail criterion between the white and black groups, it should not be used as part of the ATC selection battery. From a practical viewpoint, the exclusion of CSC 24 has essentially no impact on the multiple correlation or the predictive value for the total group or for men and women since its contribution to the prediction of the performance criterion is very small (See Tables 17 and 18, Chapter 21).

538.

Alternative Selection Procedures

During the course of this research program, a number of alternative instruments, both cognitive and non-cognitive, were examined. The 1978 studies reported in Chapter 18, of full-performance, developmental, and trainee air traffic controllers examined the relationship between non-cognitive instruments (including biographical data and personality tests) and performance criteria. Use of biographical data as an alternative selection procedure was examined in some depth.

In May 1979, the FAA proposed to the Office of Personnel Management (OPM) a demonstration project, under the Civil Service Reform Act, that addressed recruitment and selection of women and minorities in the air traffic control occupation (FAA, 1979). This proposal involved a 5-year period during which approximately half of the new ATC hires (750) in the FAA Southern region would be selected in nonregister order as a control group and the other half (750) selected in the regular manner from the OPM register to form a noncontrol goup. Candidates in the Southern region would be given a detailed multiple choice biographical questionnaire and "profiled" on relevant life experience dimensions and grouped into life profile clusters. The life profiles would then be used as an alternate means for ranking and selection of candidates for the control group. Success in ATC training would then be evaluated for both the control and noncontrol groups as a means of validating the alternative selection procedure. After review of the proposal and the biographical questionnaire, OPM pointed out a number of concerns related to job relatedness, privacy, subjectivity, and public relations for most of the questions included in the biographical form. The use was suggested of numerical scores based on "life profiles" derived from empirical data established from validity studies, and it was pointed out that the job relatedness issue presented a difficult problem because of the practical and legal requirements.

The biographical questionnaire proposed for this project was adminitered to 545 of the ATC trainees attending the FAA Academy during the period 1976-1977. Information on sex, race, OPM selection scores, ATC training laboratory composite scores and training pass-fail status was also available for this sample of 545 students. These data were provided to the Institute for Behavioral Research, University of Georgia, for analysis. Eight factors were identified through factor analysis: I. Academic factor, II. Social factor, III. Child Relationship factor, IV. Initiative factor, V. IFR/VFR Experience factor, VI. Parental Permissiveness factor, VII, Physical/Sports factor, VIII. Socio/Economic factor.

Using factor scores on these factors, the 545 trainees were clustered into six subgroups with similar life experiences. Because of the weak relationship between success-failure in the laboratory training and subgroup memberships, the eight life experience factors were used to predict the laboratory score using regression and stepwise regression analysis. The results showed that the factors were not good predictors of the laboratory

composite score which determined the pass-fail status of ATC trainees. Generally the results showed that ATC experience and tests involving mathematics and physical sciences would yield better predictions of success in ATC training. The report (Gauger, 1980) suggested that use of a sample of working controllers to establish similar life experience and group membership patterns might prove more meaningful. While this had been planned as part of the project methodology, the results of the University of Georgia study, the OPM concerns regarding validity of biographical data as a basis for selection, and cost considerations resulted in termination of this approach toward alternative selection procedures.

As pointed out in the preceding chapter, in conjunction with the analysis of the sample of 953 ATC trainees attending the FAA Academy in 1978, the FAA recognized the difficulty of selecting women and minorities from competitive OPM registers in 1968, and for that reason, established the Predevelopmental ATC program. In 1974, Executive Order 11813 provided for noncompetitive conversion of Cooperative Education students and this authority was incorporated in the recruitment and selection program for women and minorities in air traffic and other FAA occupations. These alternative selection programs have been and will continue to be a major vehicle to address the adverse impact on women and minorities as well as the Federal Equal Opportunity Recruitment Program (FEORP) requirements, established by the 1978 Civil Service Reform Act.

Table 17 shows the total number of ATC trainees hired from January 1976 through October 1980, as well as those hired from OPM register (competitive) and those hired through the Predevelopmental and Co-op programs (non-competitive) by sex and race-ethnic group.

540.

Table 17

Summary of ATC trainees hired from January 1976
through October 1980, from competitive (OPM registers)
and non-competitive (Predevelopmental and Co-op
Programs, by sex and race-ethnic group.

Group	Total Hires	Percent	Competitive Hires	Percent	Non-Competitive Hires	Percent Percent
Men	6653	84.3	6218	93.5	435	6.5
Women	1241	15.7	837	67.4	404	32.6
Total	7894	100.0	7055	89.4	839	10.6
White	6870	88.1	6442	93.8	428	6.2
Hispanic	220	2.8	134	60.9	86	39.1
Black	618	7.9	300	48.5	318	51.5
Asian	66	.9	61	92.4	5	7.6
American Indian	21	.3	18	85.7	3	14.3
Total	7795	100.0	6955	89.2	840	10.8

Chapter 23

SUMMARY OF RESEARCH ON THE EXPERIMENTAL BATTERY
RECOMMENDATIONS FOR ADOPTION AND FURTHER RESEARCH

S. B. Sells and Evan W. Pickrel

The preceding chapters in Part IV reported in some detail the re-
search that culminated in the adoption of the experimental test battery
for operational selection in October, 1981. This chapter summarizes
the principal conclusions based on that research and the recommendations
presented by the research group to implement its use.

Summary

The statistical analyses reported encompassed a large number of dif-
ferent experimental tests including the five Civil Service Commission (CSC)
tests previously used for screening of Air Traffic Controller (ATC) appli-
cants. The final experimental ATC test battery consisting of the Multi-
plex Controller Aptitude Test - MCAT, CSC 157 - Abstract Reasoning and
Letter Sequence, and CSC 24 - Arithmetic Reasoning, was derived on the
basis of multiple regression analysis. The tests were examined with re-
spect to both unweighted and weighted test score values. Only CSC 157,
of the five CSC tests, contributed to the multiple correlation (R) pre-
dicting the ATC training (performance) criterion. Weighting the tests
increased the validity of the ATC test battery.

The experimental ATC test battery, consisting of MCAT, CSC 157, and
CSC 24, was demonstrated to be a valid and statistically significant in-
strument for the preemployment screening of applicants for the ATC occu-
pation. The need for this screening is particularly important because
of: (1) the high cost of ATC training; (2) the large number of applicants
in relation to the relatively few vacancies that were filled annually, in
recent years, and (3) the fact that there is essentially no "self screen-
ing" on the part of applicants since there are no educational or specialized
experience requirements for appointment eligibility at the entry grade
levels (GS-5 or GS-7).

The use of the Occupational Knowledge Test (OKT) in conjunction with
the experimental ATC test battery increased the predictive validity of the
selection procedure. The OKT provides a more effective method of establish-
ing the applicant's knowledge relevant to the ATC occupation than the as-
sessment of related experience by means of the Rating Guide previously used
by the Civil Service Commission and its successor, the Office of Personnel
Management (OPM). Further, the use of OKT allows applicants who have ac-
quired this knowledge outside of the specific work experiences that are
given credit under the Rating Guide to earn extra credit for competitive
selection consideration.

PREVIOUS PAGE
IS BLANK

The use of the experimental ATC test battery and OKT scores to rank competitive ATC applicants for appointment consideration, and the selection of those with the highest scores first, is a valid use of the applicants' test scores, since it increases significantly the probability of success in the ATC occupation as measured by passing the Initial ATC Qualification Training Program. The addition of veterans preference points to an applicant's score on the ATC test battery is not a valid predictor of success in the ATC occupation.

Analysis of the utility of the experimental battery indicated a significant potential for substantial reduction of the overall cost of ATC training. The new battery has proven effective in the identification of applicants who passed the CSC battery, but had high fail rates in the Initial ATC Qualification training program. Operational use of the experimental battery will require the testing of increased numbers of applicants, compared to the CSC battery, in order to obtain sufficient applicants with Earned Rating scores above 85.

The establishment of the passing score on the experimental battery at the approximate mean score for an ATC applicant group is supported by the analysis of fail rates by score range groups. Setting a higher passing score (for example a transmuted score of 80) would significantly increase the difficulty in recruiting women and minorities through either competitive or non-competitive selection procedures.

The CSC battery and the experimental battery both had adverse impact on some minority groups, particularly blacks. The award of extra credit for veterans preference and ATC-related knowledge or experience to applicants who passed either test battery, also involved adverse impact on women. The experimental battery had a somewhat greater adverse impact on selection of women and blacks than did the CSC battery. Veterans preference credit was a significant factor in the adverse impact on women. The statistical evidence did not support the validity of veterans preference in relation to performance criteria, but it is required by law.

Given the evidence of adverse impact on women and some minority groups, especially blacks, analyses of test fairness were completed on the various components of the experimental test battery and the OKT. These analyses showed that the requirements of the Uniform Guidelines on Employee Selection are met for the ATC test battery comprised of MCAT and CSC 157 and for MCAT, CSC 157, and OKT extra credit points. This was true for both men and women and the white and black groups.

The Multiplex Controller Aptitude test (MCAT) is the major component in the validities obtained with the experimental battery. The analysis of fairness for MCAT, alone, included a sample of men and women as well as white and black trainees, which was proportional to the total population of ATC trainees attending the FAA Academy during the period of January 1976 through October 1980.

544.

Additional earned credit should be granted only to those applicants who achieve a score of 70 or greater on the qualifying ATC Test Battery (MCAT and CSC 157). Additional earned credit points should be granted to qualifying applicants based on OKT scores as follows:

OKT Scores (No. Right) 0-80		OKT Scores Transmuted Scale 0-100		Additional Points Earned
52-55	=	65-69	=	3
56-59	=	70-74	=	5
60-63	=	75-79	=	10
64+	=	80+	=	15

For competitive applicants, additional points based on OKT scores should be added to the scores of applicants who achieve 70 or higher on the ATC Test Battery. The total score, together with any veterans preference points, should be used to rank candidates for selection consideration on the OPM register. For non-competitive applicants, the OKT scores together with the new ATC Test Battery should be used as a basis for offering GS-7 appointments in the ATC occupation (rather than GS-5 pre-developmental appointments) to those individuals whose high scores indicate they have the aptitude and knowledge to enter the GS-7 Initial ATC Qualification training.

5. Qualified applicants should be ranked for competitive appointment consideration based on the sum of their scores on the ATC Test Battery, the OKT, and veterans preference credit, and those applicants with the highest total score (Earned Rating) should be given first consideration for selection.

When the foregoing recommendations were made, it was also recommended that the then existing OPM register of qualified applicants (approximately 4,500) should be retained, but that applicants who qualified on the basis of the new ATC Test Battery should be interspersed in their appropriate rank order with those applicants who were then on the register.

In addition, it was recommended that the six parallel forms of MCAT and the eight parallel forms of OKT, which had been developed and were used in conjunction with the validation studies, should be placed into operational use for ATC applicant screening and selection, but that the FAA, Office of Aviation Medicine, should be delegated the responsibility by OPM for continued development of additional parallel forms of MCAT and OKT, to insure against compromise of operational test forms.

Further Research with the New Battery

As explained by Pickrel in Chapter 6, the new ATC Test Battery was placed into operational use at a time of crisis, following the PATCO strike.

The addition of CSC 24 to the test battery composed of MCAT and CSC 157 revealed a significant difference in the population variance between the white and black groups, indicating that an ATC test battery comprised of MCAT, CSC 157, and CSC 24 did not meet the fairness requirements of the Uniform Guidelines for the white and black groups. The effect of these differences would be to bias the test battery somewhat in favor of the black group if a common regression line for the two groups were used. However, there were no significant differences between men and women on this test battery.

Recommendations for Operational Use of the Final Battery

The following recommendations were made by the research group, for the operational use of the final battery described above, to qualify, rate, and rank applicants for the Air Traffic Control occupation:

1. The ATC Test Battery used to qualify applicants for placement on competitive OPM registers or to qualify applicants for non-competitive appointment consideration should consist of:

   a. The two forms of the Multiplex Controller Aptitude Test
       (MCAT) with the total correct scores, weighted by a value
       of 2, and

   b. CSC Test 157, Abstract Reasoning and Letter Sequence, scored
       (R-1/4W), weighted by a value of 1.

Each form of the MCAT has 55 questions segmented into two parts; 20 minutes are allowed for Part A (27 or 28 questions) and 15 minutes for Part B (27 or 28 questions). Eight minutes are also allowed for the reading of test directions and practice problems. Therefore, a total of 43 minutes is required for administration of each MCAT form, and the total test time required for both MCAT forms is 1 hour and 26 minutes. The Abstract Reasoning and Letter Sequence test is a two-part test; each part has 25 questions. Test time for Part A is 15 minutes, and for Part B, 20 minutes. There are nine sample items with 5 minutes allowed for these practice questions. Total testing time for the recommended ATC Test Battery, including practice and test familiarization time, is 2 hours and 6 minutes.

2. The composite weighted raw score on the test battery should be transmuted by a linear transformation of the distribution so that the mean raw score of the applicant group is equated to a passing score of 70.

3. All applicants, competitive and non-competitive, who score 70 or greater should be considered eligible for appointment consideration in the Air Traffic Control occupation.

4. The Occupational Knowledge Test (OKT) should be administered to all competitive and non-competitive applicants together with the recommended ATC Test Battery. The results of the OKT should be used to grant additional earned credit in place of the previously used OPM Rating Guide.

The OKT is an 80-item test and is scored for correct answers only. Test time is 50 minutes. There are no practice items. Two minutes are allowed for reading test directions. The total test administration time, including OKT, is 2 hours and 58 minutes.

545.

At that time, the need for new controllers to replace the 11,000 who were dismissed was acute, and changes were being made in the ATCS job structure and in the Academy training curriculum.

All along, it had been intended to investigate the possibility of utilizing the new test battery differentially, for optimal assignment of qualified candidates to training for the Terminal and EnRoute options, as well as for Flight Service Station training. This research had not yet been carried out and in the press of activation of the new battery and of preparing additional alternate forms, it was put aside. The problem remains, however, along with that of the possible use of the new battery to select on the basis of qualification for radar training, which under the new system is now an advanced course at the Academy. Research on problems of differential placement should be facilitated by the increased student input at the Academy in 1982.

Despite the exceptional performance of the new test battery, the possibility exists of attaining even higher validity and therefore efficiency in prediction, by investigating additional predictors. Some possibilities that were mentioned in earlier chapters include a performance test, the Multiple Performance Task Battery, described in Chapter 4, the Dial Reading and Directional Headings tests, discussed in the EPA and Boone studies, in Chapter 18, and also new personality measures that have been developed more recently, such as those by Comrey (1970) and Jackson (1976).

The store of the MCAT, which evolved progressively from a film to a film-slide to a slide to a paper and pencil test, is extremely interesting and the extent to which Dailey and Pickrel were able to capture the dynamics of the original CODE procedure in a paper and pencil test, with the addition of aptitude items that enlarged the scope of the test, commands admiration. At the same time, it should be recognized that this tour de force in paper and pencil test construction was dictated by a policy of Civil Service Commission that prohibited the use of any tests that required apparatus to administer.

That policy may have been sound even as late as the 1970's, before the new electronic technology of computerized telecommunications engulfed the corporate and institutional scene. For the remainder of the 1980's and the future, it may finally be timely to utilize this technology for testing (excellent beginnings have been made in education and industry) and to seek a further improvement in the CODE-MCAT approach by the use of realistic, computer generated and scored simulations. The justification of paper and pencil tests and answer sheets as economical and convenient appears to be approaching a limit and there is no reason why the wave of the future should not be led by the FAA in conjunction with the implementation of the new NAS Plan, which is discussed in the remainder of this book.

547.

Part V

IMPLICATIONS FOR ATCS SELECTION OF PROJECTED
DEVELOPMENTS IN ATC SYSTEM TECHNOLOGY

Since at least the early 1970's, aware of the limitations in relation
to growth of the user population, mounting operating and maintenance costs,
and progressive obsolescence of its system hardware, FAA engineering sci-
entists have been engaged in the development of a new system, reflecting
as closely as possible the state-of-the-art in recent electronic computer
and communication technology, to replace the old. This development crys-
tallized in 1981 and was announced to the world as the New Airspace System
Plan in December, 1981, only a few months after the new selection battery
became operational.

The new system is expected to come on-line gradually, by evolutionary
steps, and to reach maturity around the year 2000. In addition to the
overview of its general architecture and major components, in Chapter 1,
Part V includes two chapters that address the human factors aspects of
the planned hardware and software and their implications for selection.
Chapter 24, by Neal A. Blake, summarizes the anticipated evolution of
air traffic control in the United States, from the system as it was in
1981 to a highly automated system, around the end of the twentieth cen-
tury. According to Mr. Blake, the new Plan, at the time of its publica-
tion, was complete mainly in concept, but much detailed implementation
was left to be achieved. His chapter gives a very clear picture of the
directions that this implementation is likely to take and forms a back-
ground for the assessment of the probable impact of the new system on the
selection of air traffic controllers.

This theme is central to the final chapter (25), by Sells and Pickrel.
Although many critical issues concerning the role of the human operator
(controller) in the future system have yet to be decided, the general
architecture and directions indicated by Blake enable some important in-
ferences, even at this early stage. These are set forth, along with the
assumptions on which they are based, as a focus for new research leading
to possible revision of the selection battery adopted for operational use
in 1981, based on the research reported in Part IV. It is estimated that
changes in the system, of sufficient importance to require changes in the
selection program, may occur as early as five years after the announcement
of the new Plan.

look like. This chapter summarizes the anticipated evolution of ATC that will take us from today's system to a highly automated system.

THE FAA AUTOMATION SYSTEM. CURRENT AND NEAR TERM

The FAA's ATC automation system is a large complex of man and machines. Control of EnRoute air traffic is provided from 25 Air Route Traffic Control Centers, which include 20 automated domestic Centers, and three automated and two manual offshore Centers. These systems interface with 182 automated and 16 manual Terminal radar approach control facilities and 447 FAA-operated Control Towers. They also interface with 3 automated and 316 manual Flight Service Stations (FSS). Future plans call for consolidating FSS operations at 61 automated facilities.

A portion of the control room in one of the centers is shown in Figure 1. Each center is organized into a number of high and low altitude control sectors where radar and procedural controllers manage the air traffic within defined airspace regions.

The present equipment of the automation system in the centers includes IBM 9020 computers and display channel equipment, which provide the automated functions and generate the information on the controller displays. In addition to the primary data channel shown at the top of Figure 1, two backup channels are currently implemented in the centers, which provide a reduced capability radar display system to be used during periods of computer outage. We are in the process of converting from use of a broadband channel, shown at the bottom, which presents a scan-converted television image of the radar data, to a much improved direct access radar channel, shown at the center, that displays aircraft and weather information to the controller, using data from the digital channel normally supplying the computer. This latter type of display more closely approaches the capability provided by the primary channel. Future enhancements to this system will permit achievement of a backup system capability that is nearly equivalent to the primary channel. It is our intention to discontinue use of the broad-band channel after the direct access radar channel is accepted for operational use.

The 9020 computer complex at one of our centers is shown in Figure 2. This equipment represents 1962 vintage technology and is physically large. Current state of the art technology will allow development of a more capable computer system approximately one-third this size.

A typical EnRoute sector suite contains radar, procedural, and assistant controller positions. Development programs are under way that are expected to replace the flight strip printers and flight strip storage bays with electronic displays of tabular flight strip data; these are discussed later.

552.

Chapter 24

THE NATIONAL AIR SPACE SYSTEM -
NOW AND AS PLANNED FOR THE YEAR 2000*

Neal A. Blake

INTRODUCTION

Over the last 25 years, the air traffic control system has evolved
from a "manual" control system based primarily on procedural separation
techniques, to an automation-aided system where many of the routine tasks
have been taken over by computers, and radar control procedures form the
basis for aircraft separation throughout most of the nation's airspace.

During the late 1960's and early 1970's, computers were installed
in the domestic EnRoute Centers and the top Terminal facilities. A cur-
rent program will result in installation of a lower capability automated
system at a number of additional radar Terminals. Our current major
system development programs will result in automation of some of the de-
cision-making functions associated with aircraft separation assurance
and metering, sequencing, and spacing of aircraft, during the 1980's.
An improved landing system and a new surveillance system with integral
ground-air-ground data link are expected to enter the system during this
same time period.

As we look forward to the next 25 years, we see a number of issues
that must be faced and decisions that must be made before the next set of
system improvements can be implemented. Some of the fundamental issues
include:

1. How far toward a fully automatic ATC system can we, and should
    we proceed?

2. What is the evolving role of the controller as the level of
    automation increases?

3. Should responsibility for a larger part of the aircraft separa-
    tion assurance function be delegated to the pilot?

FAA has recently conducted, with the user community, a New Engineering
and Development Initiatives activity which examined Policy and Technology
Choices for the future system. The user inputs resulting from this activity
have provided guidance to us in the planning and conduct of programs to
attain higher levels of automation in the ATC process. Although definition
of the future automated functions is still in a fairly early stage and no
clear picture exists, it is possible to look into the still cloudy crystal
ball and identify our best thinking on what the system of the future might

553.
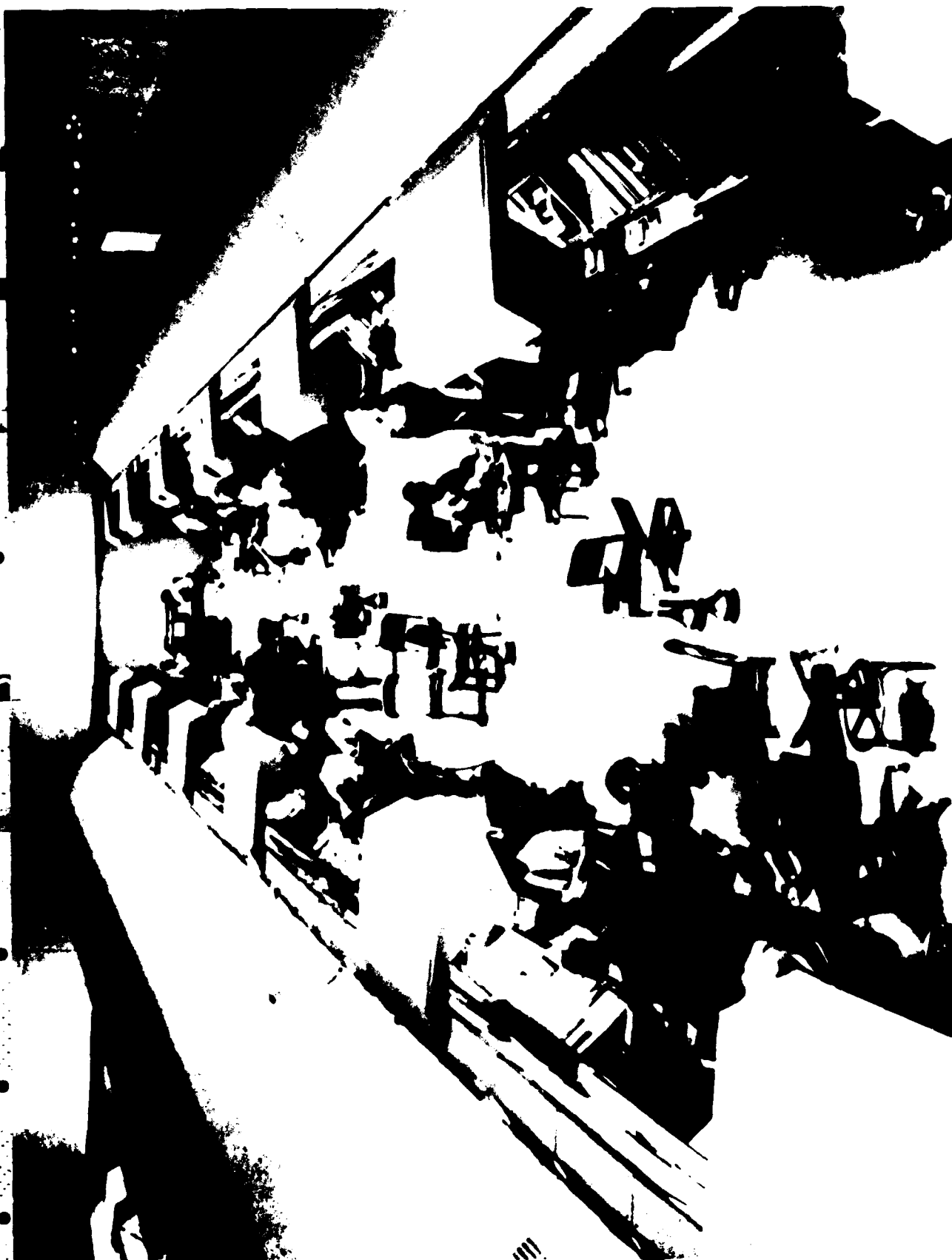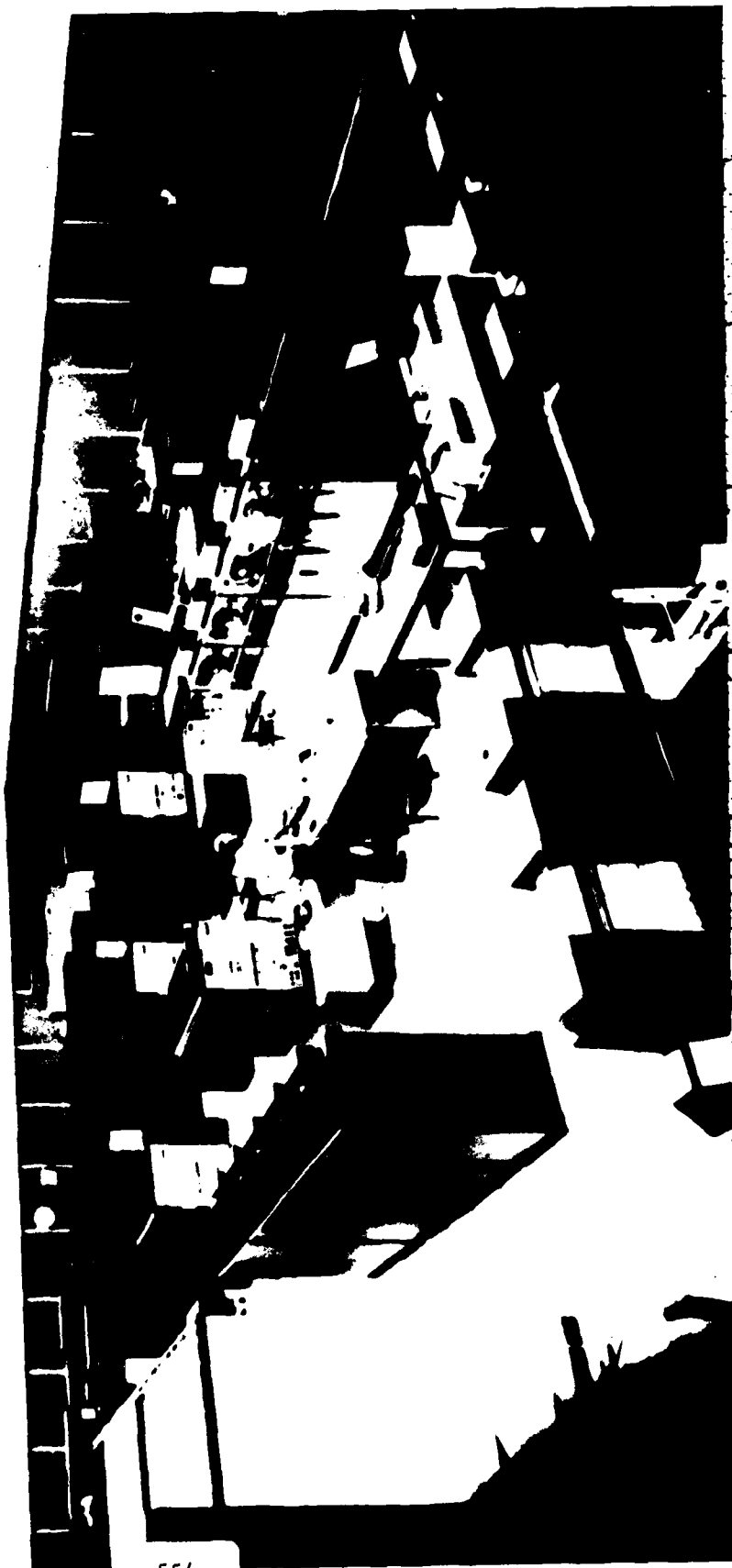
554.

The projected improvement of the Air Traffic Control System is designed to be evolutionary, not revolutionary. Future system improvements are continually under development. In the past, we have had the computer capacity to implement these new functions as they were developed. Implementation of future functions, however, will be contingent upon software improvement programs that permit "buy-back" capability. Figure 3 summarizes the functions in this evolutionary automation ladder.

Today, the primary functions performed by the EnRoute computer systems are radar and flight data processing, which provide an all digital display of aircraft track and associated information on controller radar displays and print flight progress strips at control positions.

The next functions, Conflict Alert and EnRoute Minimum Safe Altitude Warning (EMSAW), aid the controller in detecting situations where safe separation from terrain or other air traffic may be lost. The EnRoute computers also support the controller with a number of additional functions including intersector and interfacility coordination, handoff of radar identification between controllers, and generation of geographic maps and outlines of severe weather storm areas.

Functions in Development

Functions currently in the development process that are candidates for near-term implementation, subject to computer capacity availability, include the following:

1. EnRoute metering advisories to assist the controller in achieving desired aircraft flow rates in high density Terminal areas. These form the first step of an integrated flow management system.

2. Extension of the conflict alert function to warn the controller when controlled aircraft are predicted to come too close to uncontrolled aircraft operating in the same airspace.

3. Addition of conflict resolution advisories to present to the controller the range of control actions that would result in safer resolution of the conflict situation.

4. Replacement of printed flight strips with electronic tabular displays of flight data and automation of some planning activities.

5. Implementation of interfaces with the new Discrete Address Beacon System (DABS) and the Center Weather Service Unit.

555.

Automation Functions

Current System

Radar Data Processing

Flight Data Processing

Conflict Alert

En Route Minimum Safe Altitude Warning

Current System Planned Enhancements*

En Route Metering

Conflict Alert for Visual Flight Rules Intruders

Conflict Resolution Advisory

Electronic Tabular Flight Data Display Systems

Interface with the Mode S Beacon System

Interface with the Center Weather Service Unit

Replacement System Planned Enhancements

Data Link Services

Integrated Traffic Flow Management System

Automated En Route Clearance Generation

*Implementation Dependent on Computer Capacity Availability.

Major New Functions, To Be Developed

Implementation of additional major new automation functions, which are associated with automated decision-making, will not be possible until the new higher capacity, more reliable EnRoute computer system is operational. Development programs in this area are being carried on in parallel with a program to replace the EnRoute computers and will produce the system improvements needed in the post-computer replacement time period. Some of the functions in this area include:

1. Use of the Discrete Address Beacon System data link.

2. Automation of a nationwide traffic flow management service comprised of national flow control, EnRoute metering and Terminal metering and spacing functions -- all aimed at reducing airborne delays.

3. Automation of EnRoute clearance generation, to reduce planning and control workload in the centers.

Recent Improvements and Planned Additions to Current System

Some of the more significant improvements that have recently been added or are planned for addition to the system to support the controller are summarized briefly, below.

Conflict Alert warns the controller of potential separation minima violations two minutes in advance. Altitude clearances are used when inserted manually by the controller. Altitude clearance is now operational above 12,500 feet at all EnRoute centers and to the ground in many of the low altitude sectors. It can be used with primary radar targets when the controller manually inserts aircraft altitude information. In the Terminal area, implementation is nearing completion at all Automated Radar Terminal System III (ARTS III) facilities. Controller reaction to conflict alert has been positive, despite occasional complaints about false alarms. We feel that Conflict Alert is a contributor to increased safety of flight.

The Conflict Resolution Advisories function is designed to provide the EnRoute radar controller with a display of possible alternatives for the resolution of conflicts identified by the Conflict Alert function. The prime objective of the conflict resolution function is to reduce instances of system error, by reducing decision-making time in complex encounter situations. The conflict resolution function displays the range of alternatives for the resolution maneuver for the controller, who will consider factors such as desired traffic flow, severe weather, communication failures, or the presence of uncontrolled VFR aircraft, that are beyond the capabilities of the current levels of automation.

In addition to Conflict Resolution, FAA is developing two other automated systems, intended as a backup to the primary air traffic control system. These are the Automatic Traffic Advisory Service (ATAS) and the Traffic Alert and Collision Avoidance System (TCAS). ATAS is a ground-based automation function that is performed in the Discrete Address Beacon System (DABS) computer located at the radar site. TCAS is an airborne system that can protect equipped aircraft against other aircraft equipped with at least Mode C encoding transponders, both within and outside ground surveillance coverage.

Minimum Safe Altitude Warning (MSAW), is another automation feature that assists the controller in maintaining flight safety. Here the computer predicts that an aircraft is going to be below a predetermined safe altitude in the next several minutes. The flashing "LOW ALT" shown on Figure 4 warns the controller that a control action may be required.

The EnRoute metering function organizes airport arrival traffic in the EnRoute airspace by metering flights to their destination airports. Flights are scheduled for delivery to metering points for an airport at a rate that matches the acceptance rate for the airport.

The metering function determines a metering fix arrival time for each aircraft. A list of these times will be displayed to the controller for the arrival aircraft.

In order for the metered flights to meet these times, delay absorption strategies are generated to absorb any required delay. Flight progress is monitored along the approach routes, to determine the necessity for and amount of delay and, where appropriate, the optimal delay absorption strategy is developed, such as speed reduction, descent profile adjustment, or intermediate fix holding. These advisories are displayed to the EnRoute controller who is controlling the flight.

Since the inception of the air traffic control system, the method of posting flight data information to the air traffic controller has been the paper flight strip. Before the introduction of the present NAS Stage A system, flight data information was entered and updated manually by pencil on the flight strip. The present system uses electro-mechanical flight strip printers which, under computer control, print initial and updated flight data on paper strips at the sectors that will handle the flight. This system has required mounting or "stuffing" of the strips into the flight strip holders by hand, placing the holders in the desired position in the flight strip bay, updating the flight data by pencil, and entering updated information in the computer by means of a manually operated keyboard device. This is a cumbersome operation, which consumes much of the data (D) and assistant (A) controller's time.

The Electronic Tabular Display Subsystem (ETABS) development program utilizes electronic displays to replace the flight strip printers and paper

MSAW

NW123
010 32

LOW ALT
NW789
010 32

INHB
PD456
002 05

SAFETY
BUFFER

flight progress strips now in use at all EnRoute, Air Traffic Control Centers (ARTCCs). Through the use of electronic displays, processors, and touch entry devices, non-radar flight and control data will automatically be provided to the controller.

The development of ETABS forms an integral part in the design of the new controller sector suite.

The Discrete Address Beacon System (DABS) now called Mode S will have a significant impact on the controller of the future. It will provide him with more accurate and consistent surveillance information. It will provide a data link between the ground and the aircraft that opens up a variety of possibilities.

LONG TERM EVOLUTION OF THE AUTOMATED SYSTEM

The longer term automation evoluation of the system will result from the activities of the FAA Advanced Systems Engineering program. Figure 5 depicts the elements of the Advanced Systems Engineering program and their interrelationships. This discussion is limited to the most relevant elements: Automated EnRoute ATC, Integrated Flow Management, Weather Detection and Dissemination (not shown), and Data Link Applications.

The Automated EnRoute ATC Program (AERA)

Of all of FAA's automation activities, the Automated EnRoute ATC program (AERA) is likely to have the most significant impact on the controller of the future. AERA will automate most routine ATC functions that are currently performed by the controller. Since AERA will build on and incorporate the existing and near-term automation functions, it can be viewed as a logical next step in ATC evolution.

AERA will automatically perform most EnRoute planning and control processes now under the active management of controllers. This system will provide better accommodation of flexible, fuel efficient profiles; it will also increase ATC system productivity, remove many of the causes of system errors, increase ATC service availability, and reduce the potential for pilot error.

AERA will be a set of automated functions, embedded within each EnRoute facility, that will automatically plan conflict-free, fuel efficient profiles for aircraft. It will recognize aircraft conflicts 10 to 20 minutes in advance of their occurrence on the basis of current position and cleared route data, and environment conflicts, including predicted penetration of restricted airspace and severe weather areas. It will produce clearances to resolve these problems. AERA will generate routine clearances using fuel efficient profiles and direct routes, and will present these clearances to the controller. It can also deliver these messages via data link to the pilot. AERA will monitor aircraft

560.

Advanced Systems Engineering

Satellite Technology

Automated Terminal Systems

Advanced Cockpit System Engineering

Unmanned Airport

ATC Data Link Applications

561.

progress to assure separation, and will protect against system failure by providing a coast capability with backup clearances; it will be compatible with the backup TCAS system.

The AERA operational concept assumes that ground-based computers will automatically perform most of the routine ATC planning and control functions under the active management of air traffic control specialists. For IFR, EnRoute aircraft, these functions will automatically produce a plan of clearances to be issued, either automatically or manually, at appropriate times. The clearance plan, if followed, will ensure that all controlled flights will remain conflict-free, fuel-efficient, and when appropriate, metered.

Aircraft equipped with a Mode S data link, area navigation equipment, and a flight management computer will be able to take full advantage of the capabilities of the AERA system and will be able to file for and in most cases, receive precise, direct route, fuel optimal flight profiles.

In AERA, the volume of airspace within which the computer can control flight movements is known as its "control region." To avoid discontinuities in the planning control process, each AERA system begins its planning process before an aircraft enters the control region. It is anticipated that AERA control sectors can be staffed by one or two controllers and that the airspace controlled could be several times the size of current day sectors.

The productivity increases sought in AERA imply an increased traffic load for each controller, so the impact of a system failure on smooth traffic flow and on safety would be much more severe than in the current system. Thus, reliability of both hardware and software will be critical; automated backup capabilities will be needed; and human factors considerations pertaining to these backup capabilities will be important.

The status of this program is that we have formulated a detailed AERA system operational concept; and we are in the process of developing a detailed functional description of this system. In addition, a real-time AERA testbed, capable of examining the role of the controller in an automated environment, as well as demonstrating concept feasibility, is being developed. A first demonstration of the capability of AERA to plan fuel efficient flight profiles was completed in December, 1980.

Integrated Flow Management

Integrated Flow Management is a requirements definition, concept development, and integration activity. The integrated flow management concept must integrate the functions of national flow management, EnRoute metering, Terminal flow, and airport operations. This concept will use automation tools to permit the best possible integration of a variety of services and capabilities to provide fuel efficient flight with minimum delays and maximum accommodation of pilot preference.

In the Terminal area, the automation system will support Terminal planning and configuration management. It will provide vortex protection and will have the ability to provide conflict-free paths which recognize limitations imposed by weather and wind shear; and it will make use of Microwave Landing System (MLS) procedures and runway occupancy monitoring and control systems. An optimum metering, sequencing, and spacing system will ensure minimum time deviation over the threshold. We will be looking at how best to integrate these capabilities into the system and to establish the impact on ATC automation planning.

Aviation Weather Program

The overall objective of the Aviation Weather Program is to provide timely weather information to all users of the Nation's Airspace System, to improve significantly the capability for detection of hazardous weather phenomena, and to provide rapid access to the national aviation weather data base by all users. Two areas of particular interest are the development of a new generation of weather radar (called NEXRAD), and the development of Automated Weather Observation Systems, which involves the development of new weather sensors and the capability to process sensor outputs into a complete weather observation for voice and digital transmission.

Improvement in severe weather warnings involves the development of a new generation of weather radar, using doppler technology, to provide radial wind velocity and velocity spectrum width in addition to reflectivity data. Detection of hazardous weather phenomena is expected to be greatly improved through the application of doppler radar technology. A joint program was established within the Department of Commerce to develop and implement the Next Generation Radar (NEXRAD). The FAA is providing both manpower and funding resources in the NEXRAD Joint System Program Office and expects this new radar system to satisfy its requirements for the detection of hazardous weather in the EnRoute environment. It should be noted, however, what due to siting requirements, scan rates, and display update rates, the detection and display of hazardous weather phenomena in major Terminals may require a separate Terminal weather radar.

Figure 6 depicts one of the possible displays of severe weather phenomena, and shows the type of contour and storm movement information that will be made possible by NEXRAD.

Cockpit Displays of Traffic Information (CDTI)

A program to examine the use of Cockpit Displays of Traffic Information (CDTI) deserves special mention. While the technology to provide traffic information in the cockpit exists, the pilot's ability to use this information and the impact this will have on the ATC system is not fully known.

563.

Our objective is to evaluate the use of Cockpit Displays of Traffic Information for both passive monitoring and active spacing tasks so that the advantages and disadvantages of such use can be measured in terms of system safety, capacity, and efficiency in operationally realistic environments. We want to evaluate the impact of CDTI on the pilot and on the controller as well as on the traffic flow stability. Other factors to be evaluated include pilot performance in dynamic merging and spacing, display content and format, and pilot/controller work-load changes. This work is being done jointly by the FAA and NASA and is addressing general aviation and air carrier use of such displays.

Data Links

The advanced system engineering functions discussed above have two common ties. Nearly all of them require both an advanced ATC computer system and data link services. Some of the data link services that are being evaluated are shown in Figure 7. These have been grouped into several time periods with the near term functions associated with the inner ellipse. The services include:

1. Transmission of traffic and collision resolution advisories associated with ATAS and TCAS.

2. MSAW Advisories, Enhanced Terminal Information Services, EnRoute Weather, Takeoff Clearance Confirmation, and Altitude Assignment Confirmation, which are under consideration as candidates for the next time period.

3. The following are being examined as later services: Airborne Flight Plan Filing, Hazardous Precipitation Contours and Phenomena Identification Information, Metering and Spacing Clearances, ATC Instructions, Downlinking of Airborne Sensed Weather Data, Clearance Delivery, and Enhanced EnRoute Weather Information.

4. Cockpit Displays of Traffic Information and Automated EnRoute ATC Clearances, which are under consideration as post-1990 services.

We believe that the ATC automation program discussed above will provide a number of significant benefits to both users and operators of the National Airspace System. For example, system operation and maintenance costs will be reduced; user requested flight profiles will be accommodated to provide, among other things, fuel efficient flight paths; utilization of airspace and runways will be optimized; display of traffic and weather information will be improved; conflict-free flight clearances will be generated automatically; and air-ground and ground-air communications will be improved. Clearly, achievement of these benefits is dependent on the establishment of an appropriate distribution of functions between air and ground systems and personnel, and on the provision of the optimum man-machine interfaces.

Candidate Data Link Services Evolution

Advanced Automation (AERA)

Metering & Spacing Commands

Hazardous Precipitation Levels

Flight Plan Filing

Takeoff Clearance Confirmation

Altitude Assignment Confirmation

MSAW Advisory

Aircraft Separation Assurance (ASA)

Cockpit Display Traffic Information (CDTI)

ATC Instructions

Downlink Weather Information

Clearance Delivery

Enhanced En Route Weather

En Route Weather

New Computers

The benefits of increased automation do not come free or easily. New, more powerful computers, with hardware and software designed for growth and evolution, will be required. New computers will be required to support continued growth in traffic and will provide the foundation for the automation functions that have been described.

FAA is embarking on an extensive program to replace computers, to provide for growth in traffic and further automation of the ATC process. The replacement is also needed to achieve cost savings, as the cost of maintaining hardware and software systems, based on 1960's technology, is projected to increase significantly in the future.

The timetable for computer development has been to award multiple concept development contracts to industry in 1983 to begin the design of a new EnRoute computer complex. Several phases of design, development and testing have been planned, to be implemented during the late 1980's. The next generation of the controller suite for the EnRoute system will be designed and developed as part of this computer procurement.

HUMAN FACTORS IN THE TRANSITION PROCESS

The replacement of computer and display equipment in a system that must support continuous operation, without degradation in service efficiency or safety, presents a difficult transition problem. Human factors concerns will be a major factor in the transition planning. Substantial efforts will be made to minimize the impact of the transition on controllers and pilots.

First, the new system will fit into the existing communications, surveillance, and navigation environment. Second, the new system, when first installed, will look functionally identical to the old system as seen by the controller. New functions that exist in the initial replacement system will be activated gradually once transition to the new system has occurred. Third, the old system will be available as a backup for at least 90 days after the new system is placed in operation.

These transition requirements go hand in hand and will go a long way toward making the transition a smooth one for the controllers and pilots. Finally, the transition of the displays will follow the computer transition. There are two important human factors reasons for this. First, it will be much easier for the controller to adjust to a new computer system without the added problems of learning to cope with a new display. Second, this transition approach permits a logical evolution of the display systems to match the requirements of the more automated ATC system of the future.

Evolution of the EnRoute Center Sector Suite

One possible evolution of the sector suite is illustrated in Figures 8, 9, and 10. The current sector suite is shown in Figure 8, with the three control positions identified. The automation related equipment affected by the evolution includes: the radar Plan View Display (PVD), computer readout devices, data entry and select panels, and the flight strip printer.

The first major step in the evolution would be replacement of the flight strip bays at the data (D) position with electronic tabular displays. A small display, shown in the console shelf, would provide a fail safe capability for major system failure protection. This display also provides a touch entry capability for data. Note that the assistant (A) position retains the flight strip printer and some strip storage bays, to provide a system backup capability during transition.

The next step provides several new displays. The concept shown in Figure 9 has added a weather display and a new PVD to the sector suite and removed the flight strip printer and strip storage bays. This configuration is capable of supporting the Advanced EnRoute Automation (AERA) functions. Note that the present PVD is retained for backup to protect against system failures during this phase of the transition.

In the next step, the functions performed by the overhead equipment in Figure 9 have been integrated into the consoles shown in Figure 10. The present PVD console has been removed and a projection type map and auxiliary data display have taken its place. A mock-up of this concept of the future display system is available at the Technical Center and represents one possible future configuration. Other configurations are being defined and will be evaluated as part of the ongoing human factors program.

Major Human Factors Areas of Concern

Understanding the human element in ATC is an important aspect of our advanced automation programs. In conducting our current efforts we are not starting from "scratch," but building on and improving the already high performance of the current system. The focus of our current efforts is not confined to the traditional concern with location of displays and controls best suited to the physical characteristics of the human operator, although this is certainly an important area. Rather, it rests on considerations such as the following:

1. The causes and types of human error and the impact of these errors on the safety, performance and productivity of air traffic control system operations;

2. The definition of automation approaches, which assume the continued existence of human as well as machine error and strive to avoid both the occurrence and the consequences of such error;

568.

Sector Suite
Transition to ET/ABS
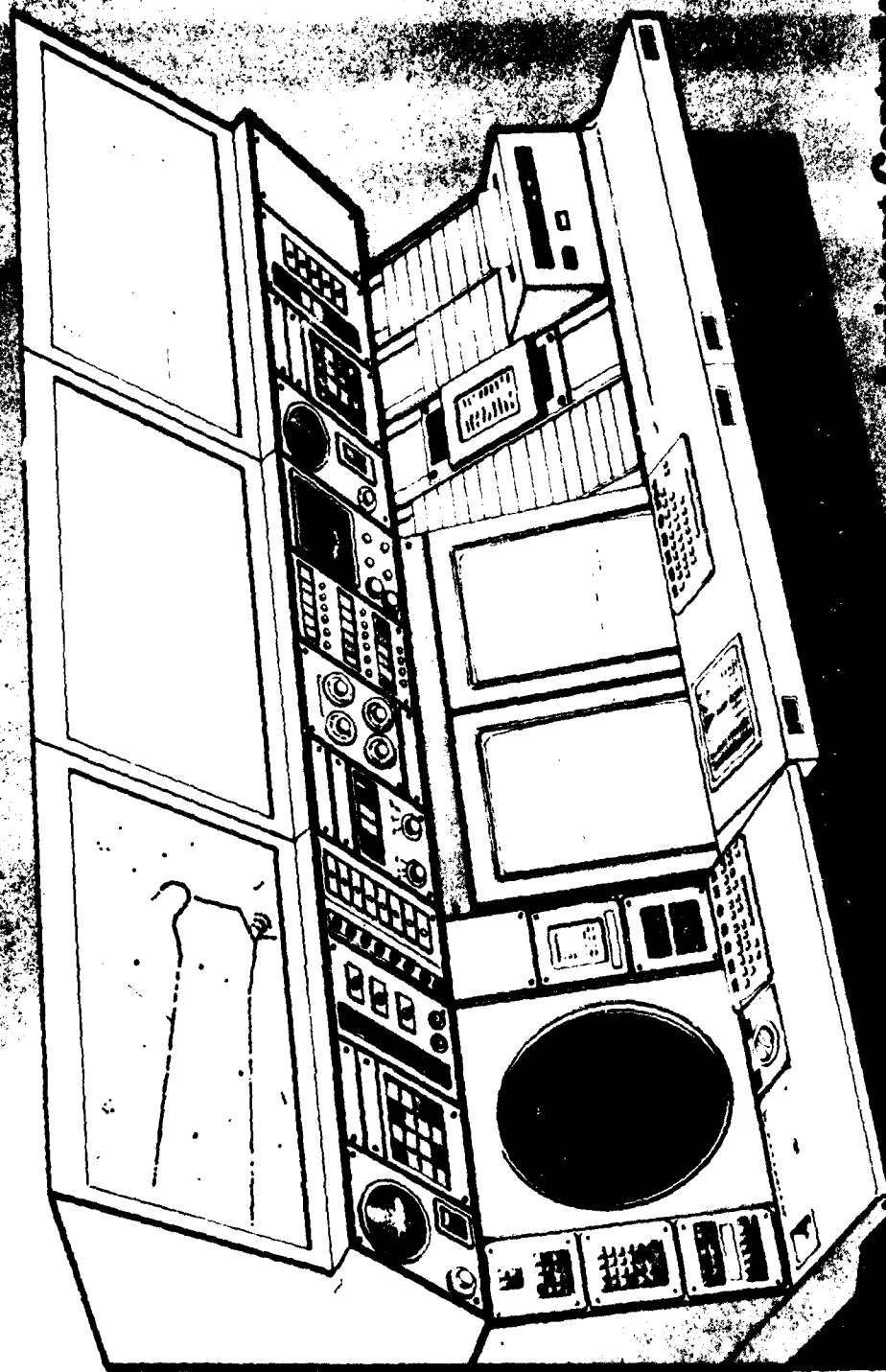
Assistant Controller

Data Controller

Radar Controller

569.

Sector Suite

Transition to ETABS and Radar

570.

1990 Sector Suite

571.

3. Assessment of the proper distribution of air traffic and
    aircraft control monitoring functions between automation
    systems and the controller and pilot;

4. Determination of the appropriate interfaces between the man
    and the machine at each step up the ladder leading to higher
    levels of automation; and

5. Determination of adequate automated, semi-automated, and
    manual system backup capabilities to permit safe continua-
    tion of system operations under a variety of conditions of
    human and machine system failure.

Human Factors Projects

Human factors considerations form an important part of the develop-
ment of new electronic data displays. These considerations span the range
from degree of automation of the flow planning process to the optimiza-
tion of data entry techniques and hardware. Touch entry and menu board
selection are of particular interest.

Controller Suite. As a part of the program for the future automa-
tion system, there is under development a set of controller suite mock-
ups, mentioned earlier, which will show several stages in the evolution
from the current to the future automated functions and associated pro-
cedures. An intra-service FAA working group was created to establish
future design requirements for the controller suite, with the specific
aims of providing design guidelines, functional descriptions, and re-
quirements for the new system. As new functions are designed and made
a part of the ATC system software, the methods for displaying data to
the controller must be carefully evaluated.

Display formats and information content. A closely associated pro-
gram is an activity to analyze the radar controller information sources,
data needs, and utilizations of currently available data, and to develop
requirements for future system display formats and information content.

Changes in the controller functions. Investigation of the controller
end of the CDTI-ATC (cockpit-controller) "interaction" loop represents
another area of investigation. This program will investigate the changes
in controller actions implied by various redistributions of the control
functions between the controller and pilot, controller impact and work-
load implications of various CDTI passive and active functions, and special
interface hardware and software design requirements needed to achieve com-
patibility between the two systems.

Color displays. Another area of investigation is the use and human
factors benefits of the introduction of color in plan view situation and
electronic tabular displays.

572.

Experimentation with higher levels of automation. Obviously, the impact of increased automation on system safety and efficiency must be demonstrated prior to implementation. Our objective, therefore, is to characterize and measure the impact of different roles for man and machine in a more automated system. As part of the AERA program, we will be using controller sector suites, such as shown in Figure 11, to assist in defining conceptual approaches to the higher levels of automation, and will make assessments of system performance at several levels of automation and associated man/machine configurations. Related activities will result in development of a systems effectiveness measurement program. In the area of ATC simulation technology and methodology, there is no currently accepted set of measures of system performance that can be utilized objectively to assess accurately the impacts of changes on the existing system. Development is under way of a system effectiveness measurement system for the evaluation of controller and system performance, to provide more objective measures of the impact of change on the system. It is also planned to develop an ATC experiment designer's handbook, that will provide objective measures to be used in assessing the impact of changes to the system.

Figure 12 lists the major activities that make up the FAA's program to address the human factors problems that have been discussed. The area of computer-aided decision-making is fundamental to the proper operation of the AERA function. The concept is based on programming into the computer a data base that will enable intelligent selection of conflict resolution and fuel efficient route profile clearances. The same data base will provide the controller greater latitude and flexibility in selecting tasks to be handed off to the computer as traffic increases and workload builds.

SUMMARY

This chapter has presented an overview of the FAA development program in the area of automation and has highlighted some of the activities related to the human factors area. As the program progresses to higher levels of automation, five areas, mentioned earlier, are considered to be of major importance. These are: (1) the impact of human errors; (2) automation approaches to avoid errors; (3) definition of the functions and role of the automation systems, controllers, and pilots; (4) the evolution of the man/machine interface; and (5) backup system requirements.

All those involved realize that the system is still in the beginning stages of the development process leading to the advanced automation system, but the process is beginning to accelerate.

574.

CPEER

CONTROLLER PERFORMANCE ENHANCEMENT AND ERROR REDUCTION

o EXAMINE CONTROLLER ROLES VS. LEVEL OF AUTOMATION

o ASSESS LEVEL OF HUMAN INVOLVEMENT

o DEFINE CONTROLLER HUMAN FACTORS PROBLEMS

o EXAMINE DISTRIBUTION OF CONTROL IMPACTS

o EVALUATE COMPUTER-AIDED DECISIONMAKING

o DEVELOP IMPROVED METHODS OF INFORMATION TRANSFER

o ESTABLISH AND VALIDATE THE RELIABILITY OF ATC SYSTEM MEASURES OF PERFORMANCE

575.

Chapter 25

ADJUSTMENT TO PROJECTED CHANGES IN THE AIR TRAFFIC
CONTROLLER ROLE AND FUNCTION

S. B. Sells and Evan W. Pickrel

By the year 2000, assuming that the NAS Plan announced in December, 1981 evolves as presently planned and on schedule, the job of the Air Traffic Controller in EnRoute centers, as presently known, will be largely preempted by the automated system. In addition, part of the present functions of controllers in Terminals will be consolidated into the EnRoute system, and to a great extent the work of the specialist in the Flight Service Station network will also be automated. However, controllers are expected to continue controlling arriving and departing traffic in the immediate vicinity of the Tower Cab, even at the busiest airports, to have advisory FSS functions and traffic control functions in the low density areas, as well as for VFR traffic throughout the system, and to man all facilities, even when fully automated. At various times during the transition period, new equipment and new procedures will be introduced and the structure of the system will be changed. The equipment, procedural, and structural changes are viewed as the key to the expected changes in the controller jobs. This chapter addresses the changes expected as inferred from information available in a dynamic planning environment.

The planned system changes have been designed to proceed in evolutionary stages, designated as (1) Near Term, to around 1985, (2) Intermediate Term, to approximately 1990, and (3) Long Term, after 1990. These boundary times are understood as approximate, serving mainly to denote specific eras marked by major events in the evolution of the system, mainly involving hardware and software acquisitions. Although there is considerable commonality, and hence redundancy in description, these can be best understood if discussed separately for the EnRoute, Terminal, and Flight Service Station systems.

REVIEW OF EQUIPMENT AND STRUCTURE BY STAGES
IN THE EVOLUTION OF THE NAS PLAN

Current Status of Systems - 1982

EnRoute systems. As of 1982 there were 20 domestic ARTCCs that handled aircraft operating under IFR rules between airport terminal areas. In addition, there were three centers, in Anchorage, Alaska, Hoholulu, Hawaii, and San Juan, Puerto Rico, that handled "offshore" traffic. A typical domestic center has been responsible for more than 100,000 square miles of airspace and thousands of miles of airways and jet routes. The geographic area of each center has generally been divided into 30 or more

577.

sectors, with a team of usually three controllers (radar, data, and assistant) responsible for each sector. The domestic centers (and sectors) have been equipped with the NAS Stage A surveillance system, consisting of a network of IBM 9020 computers that process flight data and radar data. The three offshore centers have used Sperry-Univac EnRoute Automated Radar Tracking Systems (EARTS) to perform similar, but more limited data processing functions. By 1981, these computers, developed in the mid-1960s, were approaching their limits in capacity to handle expected growth in traffic and the major enhancements to operational software required for the projected automation of the system.

The current system already incorporates some automated functions at selected facilities that are planned to be standard features of the new system. These include such features as radar beacon monitoring and handoff and various automated warning displays, such as conflict alert, en route safe altitude warning, minimum safe altitude warning, and low level wind shear advisories. The Air Traffic Control manual, dated January 21, 1982 (page 9) instructs controllers to "(a) use automation procedures in preference to nonautomation procedures when workload, communications, and nonradar separation when the situation dictates that an operational advantage will be gained." An example of the latter was given, "where vertical separation would preclude excessive vectoring." Apparently, preference decisions in these matters were left to the controller, who could suppress the automation functions according to his judgment in application of these guidelines.

At the same time that various automation systems have been entering the EnRoute controllers' environment, enabling a critical part of the system to function at least part of the time at a new technological level, the entire system has continued to function, for the remaining time, at the old level in which EnRoute controllers still initiate, issue, and when necessary, revise clearances and advisories to aircraft, transmit flight plans and estimates, receive and post weather information and NOTAMS, review, sequence, and distribute flight progress slips, load strip holders, operate interphone and radio equipment, coordinate traffic with adjacent sectors and facilities, initiate and provide Flight Assistance service, and above all, analyze the traffic picture for potential conflictions and take appropriate action.

As the new computers, sector suites, and software described below under Near Term, become available, and the planned consolidation of Centers and TRACONs proceeds, it is exp cted that both levels of functioning will continue, but that the new automation will become increasingly prevalent in the total system. Although some controllers will early on become specialized in the new procedures, there will also be others who will not encounter them until much later, while many will be required to function at both levels. This transitional situation has existed since the early 1970's and will continue, although with a shifting mix of components, until the system is fully automated, probably well after the year 2000.

According to recent discussions within the FAA, the initial implementation of AERA, to be called AERA-1, will include the controller in the customary active role, controlling traffic and sole channel of communications with pilots aloft. Good intentions apart, however, it is inevitable that the controller will no longer function autonomously. On the contrary, he or she will be "guided" by computer decisions concerning flow control and by computer determined alerts and advisories. In these circumstances, the extent to which the controller's actions may be driven by the system will probably depend on his required workload, which will be determined by sector size (and traffic density per sector) and manning of sectors.

Terminal systems. The FAA Terminal Air Traffic Control network included nearly 450 facilities, of which about 60 were closed temporarily following the controllers strike in 1981. The Terminal facilities are responsible for guiding aircraft in the immediate vicinity of an airport; many have included radar-equipped towers that provide approach and departure services up to 40-50 miles from the airport. As of 1982, 188 Terminal facilities were radar-equipped with associated computer systems that provided controllers with direct readout of the same basic flight information as the automated EnRoute systems, including the identity and altitude of all properly equipped aircraft in the Terminal area.

The two major Automated Terminal Radar Systems (ARTS) re the Sperry Univac ARTS III, at the 63 busiest airports, and the Burroughs ARTS II, at 89 medium-activity locations. A less sophisticaed automated system, the TPX 42 has been used at 35 low activity locations and these have been scheduled for early phaseout and replacement. The ARTS III has a tracking capability, whereby the computer retains a history of all aircraft equipped with a radar beacon transponder and predicts where they will be on the next sweep of the radar antenna. An upgraded version of ARTS III (ARTS III A) is capable of tracking primary radar targets as well as transponder-equipped aircraft. As a result of these capabilities, the ARTS III has been programmed to provide controllers with a Conflict Alert, which warns the controller when a potential conflict between the flight paths of any aircraft in the area is detected, and a Minimum Safe Altitude Warning (MSAW), in which the computer flashes an alert whenever an aircraft equipped with an altitude-reporting transponder is below or is predicted to go below a minimum safe altitude.

The remaining 262 Terminal facilities, located in lower density sectors, continue to utilize radio communication with departing and arriving aircraft and interfacility communications and various auxiliary systems to control departing and arriving traffic.

Flight Service systems. In 1982, there were 300 Flight Service Stations offering a broad range of preflight and inflight services, mainly to general aviation (non-airline) pilots. These services included processing of flight plans, pilot briefings on weather and safety-related matters, tracking and advising en route off-net aircraft, and assisting pilots in

distress. This has been a highly labor-intensive and costly network and automation plans are expected to reduce personnel and equipment costs, while at the same time significantly improving service to pilots, particularly in the area of more complete and timely weather reporting.

Planned Changes in the Near Term

EnRoute systems. By around 1985, the FAA plans to develop a new "host" computer to replace the 9020 system. This will be capable of utilizing the 9020 software on an interim basis with minimum modifications while providing enlarged capacity to handle the projected growth of air traffic. At the same time, two additional major developments are planned. These are (1) to design, test, and procure new sector suites equipped with distributed processing minicomputers capable of handling the functions currently performed by the display channel computers, and (2) to develop new software for the host computers that will replace the 9020 software and implement centralized processing of major new automation functions, while leaving the remaining functions to the minicomputers at the individual sector suites. Thus the Near Term will be mainly one of development of new hardware and software which are to be installed and implemented in the Intermediate Term and changes in the functions of the EnRoute controllers during the Near Term will be minimal.

The new computers are being designed to enable automated interfaces between EnRoute and Terminal TRACON facilities and the new software to implement flow control metering of high altitude traffic (above 12,500 feet), expansion of Conflict Alert to include detection of aircraft equipped with Mode C transponders when they intrude into controlled airspace, and development of conflict resolution logic programs to alert controllers (and eventually pilots) to potential violations of separation standards and recommended alternative solutions. At the same time, the mechanical, labor-intensive flight strip printers and flight data entry and printout equipment in current use will be replaced in the new sector suites by more reliable and faster electronic displays.

Also during the Near Term, it is planned to initiate reduction of the 20 domestic EnRoute Centers to 18 and consolidation of Terminal radar control facilities (TRACONS) into "hub" locations, as discussed below under Terminal Systems. Implementation of these changes must await installation of the new computers and sector suites.

Terminal systems. By the end of the Near Term, it is expected that the new computer system will be ready for installation, significant progress will have been made in the development of new system software, and the new sector suites will have been designed and tested. During the Near Term, few changes are expected in the operation of the Terminal Systems, but plans will be initiated to consolidate TRACON facilities into "hub" locations and into EnRoute centers as soon as the new computers and sector suites are available operationally. The computers provided to the hubs
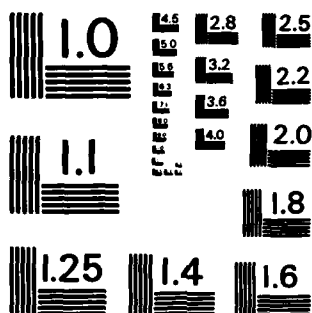
END

FILMED

DTIC

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

will be subsets of those used for the EnRoute centers and will employ common software. Sector suites at hub TRACON sites will be identical to those at EnRoute centers and will have identical data processing capability. A modular version of this sector suite will be developed for the Tower Cab, with identical processors, but with unique displays adapted to the space-limited, high intensity light environment of the Tower Cab. This sector suite will draw on EnRoute, Terminal, and FSS data bases and will satisfy the traffic control requirements for radar flight position, aircraft identity, flight data, weather data, and flow planning information.

It is expected that the new equipment will be installed at the busiest Terminals first and it is possible that some installations may be made early.

Flight Service systems. During the Near Term, the network of 300 FSS facilities is to be consolidated into 61 hubs and as of 1982 this consolidation was well under way. Contracts were awarded to purchase computerized systems for 41 of the 61 sites, with deliveries to begin in 1983. The new equipment will provide FSS specialists with rapid retrieval of weather and other flight-related data. Model 1 of the automated system will be able to display weather and aeronautical data in alphanumeric form. During the Near Term the FAA will begin implementing the more advanced Model 2 at all 61 sites. Model 2 will include a second display for weather radar, charts, and other graphics. It will also include the capability for direct access by pilots to the computer data base by remote terminals. In addition, a computer-generated voice response system will be available in certain geographical areas to provide pilots direct access by Touch-Tone telephone to limited aviation weather data, such as surface observations, terminal forecasts, and winds aloft. Replacement of low-speed teletypewriters with data terminal equipment will begin and development of improved weather sensors and weather displays will continue. Low level wind-shear equipment will be provided to additional airports and six levels of radar contouring, outlining storms, will be available on television displays for EnRoute meteorologists and specialists in automated Flight Service Stations.

Planned Developments in the Intermediate Term

EnRoute systems. The first major impact of the new system on the work of the EnRoute controller will be experienced as the old displays and display channel processors are replaced by the new sector suites and the TRACON functions are consolidated into the EnRoute centers. This will serve as a basis for future automation. The old Air Traffic Control Radar Beacon System (ATCRBS) will be replaced by the Mode S (Selective Address) system, and the completion of this replacement around 1990 will not only enable improved radar surveillance, especially in high density areas, but also the Mode S data link feature will eventually (in the Long Term) enable rapid dissemination of control and other information to specific aircraft, upon which later automation developments will depend.

581.

A complete switching and control system for voice, which will complement the capabilities of the new sector suite, is to be developed and implemented during the Intermediate Term, and this will enable the reconfiguration of sectors as needed. It was understood that a cost benefit might be realized by staff reductions as automation proceeded. Thus, as flight slips are replaced by electronically displayed flight information, and automated decision information in the form of Conflict Alerts and other advisories are routinely provided to controllers (relieving them of some cognitive processing that they previously had to accomplish without benefit of computers) the system would be expected to compensate by requiring controllers to manage increased numbers of aircraft. By this reasoning, sector size would be increased and the same number or fewer controllers per sector would be responsible for larger sectors.

Another major development during the Intermediate Term will be the upgrading of the central flow-control function; the new computer and software system will be able to project and estimate NAS congestion and delay levels and to evaluate alternate strategies for flow management, based on diverse sources of information.

It is expected that by the end of the Intermediate Term, the number of domestic centers will be reduced to 16 and only two offshore centers will operate. As consolidation proceeds, should the sectors be enlarged and staffing decreased, it is our judgment that the EnRoute controller would become increasingly dependent on the computer displays and a subtle transition would occur in the controller role. According to present plans, there will be a gradual shift in this direction. The result foreseen is that at some point, the previously predominantly independent, autonomous, deciding and controlling role of the controller will give way to one better characterized as an information link in a system in which the controller acts more as a dispatcher.

Terminal systems. The consolidation of TRACON facilities into EnRoute centers and hubs will result in removal of approach control and departure control from the Terminal ATC operation and will integrate these functions into the national flow control system. When this is complete, Terminal ATC will be restricted primarily to Tower operation for management of departure, arrival, and runway traffic at airports, but will be closely dependent on and controlled by decisions in the flow-management system.

Flight Service systems. By 1990, the consolidation of the present over 300 Flight Service Stations into the 61 automated hub facilities is expected to be completed and the Model 2 automation system should be implemented at all 61 sites. At the same time, the voice-response system capability for direct pilot access to certain weather data will be expanded and direct pilot access to the computer data base via remote computer terminals will also be implemented, providing pilot access to the data base by Touch-Tone telephone, as well. New communications switching systems

582.

will also be installed to provide direct pilot access and to expedite coordination with Flight Service Specialists. The replacement of low speed teletypewriter equipment will be completed and computer-aided direction-finder equipment will be introduced to expedite location and assistance to lost aircraft.

Major improvements in aviation weather services to pilots will be introduced during this period, such as automated weather information for pilots flying above 12,500 feet and at certain airports, using the Mode S transponder, and request-reply weather service. The weather processor at EnRoute centers will distribute current weather radar information to Flight Service specialists, Center meteorologists, Central Flow Control, and Tower Cabs at major airports.

Further Developments in the Long Term

EnRoute systems. The eventual goal of the NAS automation will be the implementation of the Automated EnRoute Air Traffic Control System (AERA) and the Integrated Flow Management (IFM) functions by the year 2000, involving a network of approximately 60 EnRoute centers and hub TRACON facilities to handle all EnRoute and approach control services in the contiguous 48 states. These systems are intended to provide IFR traffic with direct, fuel-efficient routing, flow planning and traffic management, strategic clearance delivery, and to assume full tactical control of all IFR traffic. They will use automated conflict probe and resolution, systemwide direct routing, and they are presently intended to operate with one person (controller) per sector.

To develop a nationwide system of navigation, surveillance and communications, as well as weather radar coverage, a concept called networking will be used. Mode S will replace the present secondary radar and the Microwave Landing System (MLS) will replace the current ILS. With Mode S, each aircraft will have its own beacon channel, so that it can be interrogated individually by ground-based radar and the data link feature will enable cockpit display and voice recording of automated messages in properly equipped aircraft.

At present, it is planned to automate most of the routine functions performed by air traffic controllers. Initially, the AERA system will be designed to staff all sectors with at least one controller and to place all EnRoute traffic under the active management of the sector controllers. Later, when aircraft are appropriately equipped, the communications may go directly from computer to cockpit display. In this environment, the term controller seems inappropriate and the role and functions of the person at the sector console will require careful study.

Terminal systems. By the mid-1990's, the consolidation of facilities is expected to be complete. The present 188 TRACONs will be merged into 30 newly established hub TRACONs or existing EnRoute centers. Present plans

are not definite for the residual Terminal functions in the Tower Cabs. However, it appears that they will be equipped with radar, communications, and displays, as appropriate, compatible with the new systems, so that they will be governed by the AERA and IFM decisions and have access to these data automatically. Within those constraints, however, they will have the responsibility of controlling departures and arrivals in the immediate vicinity of airports and on the ground. The major changes in these functions will involve the interface with the automated systems.

Flight Services systems. Over the long term (to the year 2000), Mode S data link coverage is planned to be extended downward from 12,500 feet to 6,000 feet, to give automated weather service to aircraft at lower altitudes. Improved weather radar data will be available as the next generation of weather radar systems, called NEXRAD, are added. To the extent that automation advances, the role of the Flight Service specialist will be increasingly directed to the service of general aviation, using the data bases available in the automated systems as a basis for providing improved services.

IMPLICATIONS OF THE CHANGES OUTLINED

In considering the projected system changes involved in the new NAS Plan, it should be understood that this was not a suddenly decided-upon, new idea that implied a brand new course for the FAA. There are indeed many new developments in this comprehensive, admirably detailed, and thoroughly integrated plan. However, to the extent that it included provision for the replacement of outmoded and obsolescent equipment, it was an overdue response to pressures experienced for some considerable time. And the incorporation in the new equipment of new (automation) capabilities enabled by progress in computer and electronic technology was a continuation of developments dating back at least 10 years that had been investigated and partially implemented by FAA systems scientists and engineers on a programmatic basis. These are exemplified by the automation programs for the 9020 computers, the Automated Terminal Radar Systems (ARTS II and ARTS III), and other computerized developments already in use in EnRoute and TRACON facilities. What the new plan accomplished, in addition to the further new features added, was integration, focus, and an announced timetable that enabled all employees, the user public (pilots and the aviation industry), and the entire world to contemplate the big picture and to gain an understanding of developments they might expect over the next 20 years.

Prior to this announcement, knowledge concerning these developments was incomplete and unevenly distributed within the FAA and little if any systematic effort had been made by the Agency to prepare its own staff and the user public for the new equipment and the new procedures entailed in its use, when changes were made. For example, a study of controllers and supervisors by the Battelle Memorial Institute (Nealey, S. M., Thornton, G. C. III, Maynard, W. S., Lindell, M. K., et al., 1975) reported

that:

"(1) People in positions of responsibility, planning, personnel, training, and other programs were not aware of the forthcoming changes, and

"(2) Few people appreciated the imminence of the changes and the necessity to begin research and program planning (on the effects of controllers). Various estimates from R & D and systems staff put the changes in the range of 6-8 years. When the (expected new) system was described to FAA personnel (controllers, managers, staff personnel, etc.) most reacted either that such things were impossible, they were too expensive, not practical, or were many years away. When told that the changes were scheduled for implementation within 6-8 years, most controllers simply did not believe it would happen."

The failure to inform and the absence of standardization and systematic training in the operation of new computerized equipment appears to reflect a hardware-oriented bias in the agency, characterized by preoccution with technical and engineering aspects and neglect of human aspects, particularly human engineering (the interface of the human operator with the system of which he or she is a part) and personnel management. The Battelle report noted widespread (but not universal) resistance to changes that had been introduced that involved "less voice communication (with pilots and other controllers) and handwritten work (flight strips) but more manual keyboard tasks; ...less burden of soliciting and remembering aircraft identification and altitude information but more necessity to remember computer format rules." It also described widespread personnel problems that unquestionably culminated in the 1981 controller strike.

The complex problems of personnel management in the FAA were addressed by a task force headed by Lawrence M. Jones that was commissioned in 1981 by the new FAA Administrator, J. Lynn Helms (Jones, Fuller, and Bowers, Note 1) and are beyond the scope of this book, although they reflect the same general bias of "benign neglect" of the human problems in air traffic control that existed previously in the program and that appears to be continuing in the human engineering area. The systems that have been designed and detailed in this and earlier chapters have focused on work output, but have virtually ignored the changes in information processing skills required for controller performance and the impact of these changes on controller motivation, job satisfaction, and performance. Even in 1975, Nealey et al. raised such questions as, The controller will be working more with a computer, but in what ways?, What will the job look like to the controller?, Can current controllers be trained to do it?, and Can anyone be trained to do it?

These and related questions appear not to have been addressed, but nevertheless the issues raised by them have been realized, as in the discussion by Blake in Chapter 24 concerning the plan to establish a real-time

test bed to study the role of the controller in an automated environment. At times it has seemed that they have been taken for granted on the assumption that people always adapt.

The general impact of automation on the EnRoute and TRACON controllers, in particular, was characterized by Nealey et al. as involving changes:

from an active role to a more passive role,
from the function of monitoring, and hence
from a proactive posture to a reactive posture;
from primarily verbal communication activity to
manual activity, and
from a goal of "control to maximize" to one of
"monitor to standard."

Our present examination of the subtle changes in the job implied by the automation of decision functions, the possibility of enlargement of sectors accompanied at the same time by reduction of sector controller staff, which would imply drastic increases in sector traffic load, reduced interaction with pilots and other controllers, and the implied change from active decider and controller to passive relayer of information, is in agreement with this characterization.

SUGGESTED NEW RESEARCH RELATED TO CHANGES ANTICIPATED

The discussion thus far has suggested that controller adaptation to increasing automation of the system may well involve attitudinal and fundamental personality issues as well as aptitudes and skills related to performance under the changed conditions. At least three related lines of research are indicated to provide a knowledge base leading to the development of new controller selection instruments for the future. These are discussed below under the following headings: 1. motivational studies, 2. predictive studies, and 3. simulation studies.

Motivational Studies

In view of the findings of Nealey et al. (1975) of some significant amount of controller resistance to the introduction of automation procedures, and the policy of FAA concerning their use, as quoted from the Controllers' Manual (1982), above, it appears timely to obtain hard data concerning controller use of the automation procedures in the current system environment. Such information would be most useful if collected systematically for statistically adequate subsamples of both EnRoute and Terminal Controllers whose work assignments represent various combinations of specific positions, technological patterns, and types of location or facility, such as the following:

586.

Option	Mean Traffic Density at Facility	Type of Equipment Available	Position
EnRoute	High	Automation available	Radar
Terminal	Medium	Automation not available	Data
	Low		Assistant

Individual information concerning test scores, training, experience, sex, minority-majority status, and other variables could be considered at the same time.

For those subsamples for which automation equipment is currently available, observational and interview data could be collected concerning the level of use of automation procedures in order to determine the extent to which use is related to traffic density (and work load), position (and experience level), training (including specific training to use the automated procedures), use level at facility, supervision, test scores, Academy grades, and other variables. Obstacles to use of automation procedures could be determined and solutions sought. For all groups, information should be obtained concerning knowledge and attitudes about automation that could be related to the same variables.

Such research supplemented by detailed task analysis studies of the jobs performed by each of the subsamples identified would answer questions concerning the extent of use of automation procedures where they should be used, and whether nonuse is primarily a selection, training-indoctrination, supervision, or human factors problem, or some combination of these. It might provide guidelines for overcoming resistance to change and negative attitudes toward the new equipment and procedures.

Predictive Studies

Ordinarily, when a new employee selection system is adopted operationally, data are collected on a continuing basis for periodic review of its effectiveness. This enables assessment of the validity and utility of the system over time and may provide indications of need to consider revision, particularly when changes are known or suspected to be occurring in the applicant population or in any significant aspects of the job or its operational environment. Undoubtedly, the FAA will monitor the new battery for several years, perhaps throughout the Near Term.

Considering the fact that the new battery was validated for combined samples of EnRoute and Terminal controller trainees, and against training level criteria, it appears desirable to focus further analyses of validity, not only on new classes completing Academy training, taking account of any curriculum changes and other significant events in the training environment that may occur, but also on classes completing radar training and on former students selected by the new battery as they proceed through developmental

587.

and journeyman stages, using such post-training criteria as may be available. Grouping these former students into the subsamples described above, the post-training studies would provide a valuable assessment of the predictive validity of the new battery for groups performing different functions relevant to the transition to the new technology. At the same time, the detailed task analysis of the jobs performed by the subsamples analyzed might lead to the generation of hypotheses concerning new predictors that might be tested simultaneously with the predictive studies of the new battery, in concurrent designs.

Simulation Studies

Human factors studies in a real-time test bed are much needed to determine precisely what functions the human operator can and should be expected to perform in the new NAS system, how large a traffic load a single operator can preside over as a sector monitor-dispatcher in an automated system, and for how long at a sitting.

Such studies are urgently needed since the present basis of selection and training of air traffic controllers, as reflected by the Multiplex Controller Aptitude Test, the key instrument in the new selection battery, and the present training curriculum, is premised on the rationale that the core of controller proficiency is efficiency in the detection and prevention of conflictions. The direction of change in the controller job is away from analysis of potential conflictions as the critical task to be performed. This is to be done by the computer, while other tasks, including monitoring the computer displays, inputting information to the computer, and reacting to computer directions are to be emphasized for the operator. The original NAS Plan (1981) suggested that as soon as flight strips are replaced by automatic displays and conflict alerts and other warnings are provided, sectors would be enlarged to compensate for the processing time that controllers previously required to perform these tracking and decision functions. However, later statements indicated that such changes would be studied, and approached very cautiously. In our judgment, such study and caution are extremely important, since situations might readily occur in which the activity on the display boards would require the full attention of the controller and would preclude the possibility that the controller might have enough information about the current status of the traffic to question the computer reliably and override its decisions, except perhaps in unusual cases. It appears even more doubtful that the controller would be able to "take over" in the event of system failure, as suggested by several writers.

One consequence of the type of situation described appears to be that the field of the controllers attention would necessarily shift from visualization of ongoing traffic in real life terms to preoccupation with the computer generated information on the display board. This may not be necessarily bad, in terms of system effectiveness, but it does imply a requirement for different aptitudes and perhaps a different person profile than that which has been found to be optimal in the present system.

588.

A profile of the present controller is given in Dailey's chapter
(Chapter 7) and this was amplified by some observations in the Nealey
et al. (1975) report. According to Dailey:

"The central skill of the controller seems to be the ability
to respond to a variety of quantitative inputs about several air-
craft simultaneously and to form a continuously changing mental
picture to be used as the basis for planning and controlling the
courses of the aircraft." (page 13)

"In preparation for these decision-making functions, the
air traffic controller must master a vast library of manuals,
maps, regulations, letters of agreement, and similar publica-
tions that guide the planning and decision making. These cover
such areas as weather, communications, navigation, geography,
flight service, aeronautics, and flight regulations. All of
these are constantly changing, and the controller must keep up
to date on them continuously. The controller must also be adept
at reading maps, charts, and tables and performing mental arith-
metic on the data read. The combination of these mental demands
requires that the controller be a "perceptual-discrimination
athlete." (page 14)

"The air traffic control system is a man-machine system
which places primary reliance on the human component, and the
human operator is often the limiting factor in system output.
The air traffic control system violates the first tenet of man-
machine design -- that is -- 'to design the system so the average
man with an average amount of effort can carry out his part in
the system.' Unfortunately, there seems to be no feasible way
to do this and so the system requires a controller who is highly
selected and highly trained. The controller bears a heavy load
on his memory and his ability to keep many things in mind while
arriving at well-reasoned solutions to complex problems under
conditions of stress. Controlling air traffic is not necessarily
a stressful activity, but in situations with a high traffic den-
sity, the job of the air traffic control specialist becomes
stressful because it requires pushing man's capability to its
limit to maintain continuous peak performance." (page 24)

This profile of high aptitude for spatial visualization, coupled with
high learning ability and both short-term and long-term memory, facility in
mental arithmetic, quantitative reasoning, and verbal facility has con-
tinued to describe the effective air controller into the early 1980's and
is expected to be valid throughout the Near Term of the new Plan. This is
expected to be true despite some continuing automation of information pre-
sentation to the controller, particularly at the busiest Centers. It may
lose relevance during the Intermediate Term, when the new computers and
sector suites are installed and the integration of EnRoute centers and
TRACONs, and enlargement of sectors are expected to take place.

If the installation of the new system is accompanied by the drastic reduction of staff mentioned in the system plan (namely, one controller per sector) and should sector size be commensurately increased, then, at some point, around 1990, the transition from active, deciding controller to passive dispatcher, controlled by the computer could be expected to be completed, unless problems incountered along the way should force some changes in the system. Both Dailey and Nealey et al. have emphasized the importance of engineering the automation system to provide an active role for the controller. Under the planned system, this might be accomplished by restricting sector enlargement to a size that would enable the controller to utilize the automated information while at the same time having the opportunity to keep track of the traffic mentally. Determination of this limit is a research problem that could be addressed in the real-time test bed studies mentioned earlier.

OVERVIEW OF THE TOTAL SYSTEM

Our preoccupation with the anticipated automation of the NAS system should not cause us to lose sight of the total system, which includes general aviation and VFR traffic as well as commercial airliners and a small portion of the general aviation fleet that will be equipped with the transponders and cockpit display gear that airliners are expected to carry. In view of the developments outlined, it appears reasonable to expect that the diversity of function among air traffic controllers will be considerably greater in the automated era. And this may be true even if EnRoute and Terminal procedures are highly standardized in the automated system, in contrast to the largely unstandardized situation that presently prevails. It will still be necessary to assign clearances, monitor departures and arrivals, and provide the full range of flight services and there will continue to be the less dense and less highly traveled sectors. The new Plan is so complex that it was not feasible to address this part of the total system, but this will be necessary by the end of the Near Term, when the impact of the system changes will begin to be felt.

As the Near Term progresses, it appears that the FAA will have around five years (assuming no delays) to evaluate the impact of the developing system on the air traffic controller and either to modify the system to accommodate to the characteristics of the human operator or to investigate the changed requirements for the operator (to change the selection program if a feasible new role is defined), or both. Compared to the time devoted to the research on controller performance and selection described in this book, this alarm comes quite late and should be considered as one of major urgency to assure success of the program.

REFERENCE NOTE

1. Jones, Lawrence M., Fuller, Stephen H., and Bowers, David G. Management and employee relations within the Federal Aviation Administration. Task Force Report, Contract No. DTFA01-82-C-3006. Federal Aviation Administration, 1982.

REFERENCES

Adams, J. A. Some considerations in the design and use of dynamic flight simulators. Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, AFPTRC-TN-57-21, April 1957.

Adkins, D. C. Construction and analysis of achievement tests. Washington, D. C.: U.S. Government Printing Office, 1947.

Air Traffic Services. Air traffic control. Washington, D. C.: U.S. Department of Transportation, Federal Highway Administration, 1982.

Allport, G. W., & Odbert, H. S. Trait-names: A psycho-lexical study. Psychological Monographs, 1936, 47 (1, Whole No. 211).

Anastasi, A. Psychological testing (3rd ed.). New York: Macmillan, 1968.

Bale, R. M., & Ambler, R. K. Application of college and flight background questionnaires as supplementary non-cognitive measures for use in the selection of student naval aviators. Aerospace Medicine, 1971, 42, 1178-1181.

Bartanowicz, R. S. The Armed Services Vocational Aptitude Battery and the Federal Aviation Administrative's Controller Decision Evaluation and Multiplex Controller Aptitude Tests as Predictors of Success in the USAF Air Traffic Control Field, Dissertation presented to the faculty of the Graduate School of Arts and Sciences Univ. of Denver, Air Command Staff College, 1979.

Barzun, J. The wasteland of American education. New York Times Review of Books, 1981, 28, 34-36.

Bolton, B. Evidence for the 16PF primary and secondary factors. Multivariate Experimental Clinical Research, 1977, 3, 1-15.

Boone, J. O. The use of path models to study a precareer air traffic control training program. Aviation, Space, and Environmental Medicine, 1978, 49, 1203-1211.

Boone, J. O. Toward the development of a new selection battery for air traffic control specialists. Washington, D.C.: Federal Aviation Administration, FAA-AM-79-21, 1979. (a)

Boone, J. O. A statistical procedure for eliminating extreme, deviant scores from the longitudinal air traffic control data base. In J. O. Boone & M.A. Lewis (Eds.), The selection of air traffic control specialists: Two studies demonstrating methods to insure an accurate validity coefficient for selection devices. Washington, D.C.: Federal Aviation Administration, FAA-Am-79-14, 1979. (b)

Boone, J.O. Toward the development of a new aptitude selection battery for air traffic control specialists. Aviation Space and Environmental Medicine, 1980, 51, 694-699.

Boone, J. O. A generic model for evaluation of the Federal Aviation Administration air traffic control specialist training programs. Washington, D. C.: Federal Aviation Administration, FAA-AM-82-2, 1982.

Boone, J. O., & Lewis, M. A. The development of the ATC Selection Battery: A new procedure to make maximum use of available information when correcting correlations for restriction in range due to selection. Washington, D.C.: Federal Aviation Administration, FAA-AM-78-36, 1978.

Boone, J. O., & Lewis, M. A. A demonstration of possible effects of recruitment and selection procedures on correcting the validity coefficient for restriction in range. Psychological Reports, 1980, 46, 927-930.

Boone, J. O., Van Buskirk, L., & Steen, J. The Federal Aviation Administration's Radar Training Facility and employee selection and training. Washington, D.C.: Federal Aviation Administration, FAA-AM-80-15, 1980.

Booze, C. J. The morbidity experience of air traffic control personnel, 1967-1977. Washington, D.C.: Federal Aviation Administration, FAA-AM-78-21, 1978.

Boyle, D. Air traffic control automation. Interavia, 1975, 5, 491-497, 531-534.

Brokaw, L. D. Selection measures for air traffic control training. Lackland Air Force Base, TX: Personnel Laboratory, Air Force Personnel and Training Research Center, Technical Memorandum PL-TM-57-14, July 1957.

Brokaw, L. D. School and job validation measures for air traffic control training. Lackland AFB, TX: Wright Air Development Center, United States Air Force, WADC-TN-59-39, 1959.

Brokaw, L. D. Historical overview of research and development of air traffic controller selection. Presented at Military Testing Association, San Diego, CA, October 15-19, 1979.

Buchanan, J. C., David, S. O., & Dunnette, M. D. Behavioral reliability program for the nuclear industry. Washington, D.C.: U. S. Nuclear Regulatory Commission, Report NUREG/CR-2076, 1981.

Buckley, E. P. Development of a performance criterion for enroute air traffic control personnel research through air traffic contral simulation: Experiment 1 - parallel form development. Atlantic City, NJ: National Aviation Facilities Experimental Center, Federal Aviation Administration, FAA-RD-75-186, 1976.

593.

Buckley, E. P., & Beebe, T. The development of a motion picture measure-ment for aptitude for air traffic control. Atlantic City, NJ: National Aviation Facilities Experimental Center, FAA-RD-71-106, 1972.

Buckley, E. P., DeBaryske, B. D., Hitchner, N., & Koher, P. Methods and Measurement in Real-Time Air Traffic Control System Simulation. At-lantic City, NJ: Technical Center, Federal Aviation Administration, DOT/FAA/CT-83/26, 1983.

Buckley, E. P., House, K., & Rood, R. Development of a performance cri-terion for air traffic control personnel research through air traffic control simulation. Washington, D.C.: Federal Aviation Administra-tion, FAA-RD-78-71, 1978.

Buckley, E. P., O'Connor, W. F., & Beebe, T. A comparative analysis of individual and system performance indices for the air traffic control system. Atlantic City, NJ: National Aviation Facilities Experimental Center, Federal Aviation Administration, NA-69-40, 1969.

Buckley, E. P., & Rood, R. H. CPM PROBE Experiment on performance informa-tion feedback. Atlantic City, NJ: National Aviation Facilities Ex-perimental Center, Federal Aviation Administration, NA-77-18-LR, 1977.

Burkhardt, R. The Federal Aviation Administration. New York: Praeger, 1967.

Butcher, J. N. Personality assessment: Problems and perspectives. In J. N. Butcher (Ed.) Objective personality assessment. New York: Academic Press, 1972.

Butcher, J. N. Use of the MMPI in personnel selection. In J. N. Butcher (Ed.) New developments in the use of the MMPI. Minneapolis: Uni-versity of Minnesota Press, 1979.

Campbell, J. B., & Chun, K-T. Inter-inventory predictability and content overlap of the 16PF and the CPI. Applied Psychological Measurement, 1977, 1, 51-63.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. Managerial behavior, performance, and effectiveness. New York: McGraw-Hill, 1970.

Campbell, J. P. & Pritchard, R. D. Motivation theory in industrial and or-ganizational psychology. In M. D. Dunnette (Ed.) Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Carver, J. Education in America: Its decline and possible fall. Humanist, 981, 41, 47-49, 56.

Cattell, R. B. Personality, a systematic theoretical and factual study. New York: McGraw-Hill, 1950.

Cattell, R. B. The description and measurement of personality. New York: World, 1946.

Cattell, R. B. Personality and motivation: Structure and measurement. New York: World, 1956.

Cattell, R. B. The 16PF and basic personality structure: A reply to Eysenk. Journal of Behavioral Science, 1972, 1, 169-187.

Cattell, R. B. Personality and Mood by questionnaire. San Francisco: Jossey-Bass, 1973.

Cattell, R. B., & Eber, H. W. The sixteen personality factor questionnaire. Champaign, IL: Institute for Personality and Ability Testing, 1962.

Cattell, R. B., Eber, H. W., & Delhees, K. H. A large sample cross validation of the 16PF with some clinical implications. Multivariate Behavioral Research, 1968, 3, 107-132.

Cattell, R. B., Stice, G. F., & Eber, H. W. The 16 personality factor questionnaire. Champaign, IL: Institute for Personality and Ability Testing, 1949.

Chiles, W. B., Alluisi, E. A., & Adams, O. S. Work schedules and performance during confinement. Human Factors, 1968, 10, 143-196.

Chiles, W. D., Jennings, E. E., & West, G. Multiple task performance as a predictor of the potential of air traffic controller trainees. Washington, D.C.: Federal Aviation Administration, FM-AM-72-5, 1972.

Chiles, W. D., & West, G. Multiple task performance as a predictor of the potential of air traffic controller trainees: A follow-up study. Washington, D.C.: Federal Aviation Administration, FAA-AM-74-10, 1974.

Circular No. A-46. Standards and guidelines for federal statistics. Washington, D.C.: Office of Management and Budget, Executive Office of the President, May, 1974.

Cobb, B. B. Problems in air traffic management: II. Prediction of success in air traffic controller school. Aerospace Medicine, 1962, 33, 702-713. See also: Problems in air traffic management: II. Prediction of success in air traffic controller school. Washington, D.C.: Federal Aviation Administration, FAA-AM-62-2, 1962.

Cobb, B. B. Problems in air traffic management: V. Identification and potential of aptitude test measures for selection of tower air traffic controller trainees. Aerospace Medicine, 1964, 35, 1019-1027.

Cobb, B. B. Problems in air traffic management: V. Identification and potential of aptitude test measures for selection of tower air traffic controller trainees. Washington, D.C.: Federal Aviation Administration, FAA-AM-65-19, 1965.

Cobb, B. B. The relationships between chronological age, length of experience, and job performance ratings of air route traffic control specialists. Washington, D.C.: Federal Aviation Administration, FAA-AM-67-1, 1967.

Cobb, B. B. Relationships among chronological age, length of experience, and job performance ratings of air route traffic control specialists. Aerospace Medicine, 1968, 39, 119-124.(a)

Cobb, B. B. A comparative study of air traffic trainee aptitude test measures involving Navy, Marine Corps and FAA controllers. Washington, D.C.: Federal Aviation Administration, FAA-AM-68-14, 1968.(b)

Cobb, B. B. Air traffic aptitude test measures of military and FAA controller trainees. Washington, D.C.: Federal Aviation Administration, FAA-AM-71-40, 1971.

Cobb, B. B., Lay, C. D., & Bourdet, N. M. The relationship between age and aptitude test measures of advanced-level air traffic control trainees. Washington, D.C.: Federal Aviation Administration, FAA-AM-70-14, 1970.

Cobb, B. B., Lay, C. D., & Bourdet, N. M. The relationship between chronological age and aptitude test measures of advanced-level air traffic control trainees. Washington, D.C.: FAA Office of Aviation Medicine Report No. AM-71-36, 1971.

Cobb, B. B., & Mathews, J. J. Proposed new test for aptitude screening of air traffic controller applicants. Aerospace Medicine, 1973, 44, 184-189.

Cobb, B. B., Mathews, J. J., & Lay, C. D. A comparative study of female and male air traffic controller trainees. Washington, D.C.: Federal Aviation Administration, FAA-AM-72-22, 1972.

Cobb, B. B., Mathews, J. J., & Nelson, P. L. Attrition-retention rates of air traffic control trainees recruited during 1960-63 and 1968-70. Washington, D.C.: Federal Aviation Administration, FAA-AM-72-33, 1972.

Cobb, B. B., & Nelson, P. L. Aircraft pilot and other preemployment experience as factors in the selection of air traffic controller trainees. Washington, D.C.: Federal Aviation Administration, FAA-AM-74-8, 1974.

Cobb, B. B., Nelson, P. L., & Mathews, J. J. The relationships of age and ATC experience to job performance ratings of terminal area traffic controllers. Washington, D.C.: Federal Aviation Administration, FAA-AM-73-7, 1973.

Cobb, B. B., Nelson, P. L., & Mathews, J. J. Relationships between age, ATC experience, and job ratings of terminal area traffic controllers. Aerospace Medicine, 1974, 45, 56-60.

Cobb, B. B., Young, C. L., & Rizzuti, B. L. Education as a factor in the selection of air traffic controller trainees. Washington, D.C.: Federal Aviation Administration, FAA-AM-76-6, 1976.

Cohen, J. Statistical power analysis for behavioral science. New York: Academy Press, 1969.

Collins, W. E., Boone, J. O., & VanDeventer, A. D. The selection of air traffic control specialists: I. History and review of contributions by the Civil Aeromedical Institute. Washington, D.C.: Federal Aviation Administration, FAA-AM-80-7, 1980.

Collins, W. E., Boone, J. O., & VanDeventer, A. D., (Eds.). The selection of air traffic control specialists: History and review of contributions by the Civil Aeromedical Institute, 1960-80. Aviation, Space, and Environmental Medicine, 1981, 52, 217-240.

Colmen, J. G. Review and evaluation of present system for selection of air traffic controllers. Washington, D.C.: Education and Public Affairs, Inc., FAA Contract DOT-FA-70WA-2371, Report for Phase I, Task I., 1970.

Colmen, J. G. Validity of the Cattell 16 Personality Factor Questionnaire and other "non-cognitive" tests for selection and placement of air traffic control specialists. Washington, D.C.: Education and Public Affairs, FAA Contract DOT-FA-75WA-3646, 1977.

Committee on Government Operations, Twelfth Report. Selection and training of FAA air traffic controllers. Washington, D.C.: U.S. Government Printing Office, January 1976.

Corson, J. J., Chairman. Air Traffic Controller Committee Report: The career of the air traffic controller - A course of action. Washington, D.C.: Federal Aviation Administration, 1970.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Crowder, N. A. _Proficiency of Q-24 radar mechanics: V. Level of trouble-shooting performance observed._ Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, AFPTRC-TR-54-102, 1954.

Dailey, J. T. _Development of an Occupational Knowledge Test for ATCS._ FAA Office of Aviation Medicine Report (Unpublished), April 1971.

Dailey, J. T., & Moore, J. I. Criterion development for air traffic controller training. In _Selection and evaluation of air traffic controllers_, Proceedings of the 21st Annual Military Testing Association Conference, San Diego, California, 1979.

Dailey, J. T., & Pickrel, E. W. Federal Aviation Administration behavioral research program for defense against hijackers. _Aviation, Space, and Environmental Medicine_, 1975, _46_, 423-427. See Also: Some psychological contributions to defenses against hijackers. _American Psychologist_, 1975, _30_, 161-165.

Dailey, J. T. & Pickrel, E. W. _Development of new selection tests for air traffic controllers._ Washington, D.C.: Federal Aviation Administration, FAA-AM-77-25, 1977.

Dailey, J. T., & Shaycroft, M. F. _Types of tests in Project Talent OE-2501._ U.S. Cooperative Research, Monograph No. 9, 1961.

Danoher, J. W. Human error in ATC systems operations. _Human Factors_, 1980, _22_, 535-546.

Davis, C. G., Kerle, R. H., Silvestro, A. W., & Wallace, W. H. _The air traffic control training program as viewed by training supervisors._ Washington, D.C.: Federal Aviation Agency, Technical Report No. 33, Project O, Bureau of Research and Development, 1960.

Davis, C. G., Kerle, R. H., Silvestro, A. W., & Wallace, W. H. _Identification of training requirements in air traffic control._ Washington, D.C.: Federal Aviation Agency, Technical Report No. 36, Project O, Bureau of Research and Development, 1960.

DuBois, P. H., Loevinger, J., & Gleser, G. C. _The construction of homogeneous keys for a biographical inventory._ San Antonio, TX: Lackland Air Force Based, Air Training Command Human Resources Research Center, Research Bulletin 52-18, 1952.

DeYoung, G. E. Standards of decision regarding personality factors in questionnaires. _Canadian Journal of Behavioral Science_, 1972, _4_, 253-255.

Ebel, R. L. Estimation of the reliability of ratings, _Psychometrika_, 1951, _16_, 407-424.

598.

Eysenck, H. J.   On the choice of personality tests for research and pre-
    diction.   Journal of Behavioral Science, 1971, 1, 85-89.

Eysenck, H. J.   Primaries or second-order factors: A critical considera-
    tion of Cattell's 16PF battery.   British Journal of Social and Clinical
    Psychology, 1972, 11, 265-269.

Eysenck, J. J., White, P. O., & Soueif, M. I.   Factors in the Cattell per-
    sonality inventory.   In H. J. Eysenck & S. B. Eysenck (Eds.), Person-
    ality structure and measurement.   San Diego: Robert R. Knapp, 1969.

FAA Task Force Report: Air traffic controller selection and retention.
    Washington, D.C.: Federal Aviation Administration, 1975.

Federal Aviation Administration.   National airspace system plan, facilities,
    equipment, and associated development.   Washington, D.C.: Federal
    Aviation Administration, 1981.

Finkelman, J. M., & Kirschner, C.   An information-processing interpretation
    of air traffic control stress.   Human Factors, 1980, 22, 561-568.

Flanagan, J. C.   Scientific development of the use of human resources:
    Progress in the Army Air Forces.   Science, 1947, 105, 57-60.

Flight Safety Foundation.   A safety appraisal of the Air Traffic Control
    System.   Arlington, VA: Flight Safety Foundation, Inc., DTFA01-81-C-
    10109, 1982.

Frank, F. D., Lindley, B. S., & Cohen, R. A.   Standards for psychological
    assessment of nuclear facility personnel.   Washington, D.C.: U.S.
    Nuclear Regulatory Commission, Report NUREG/CR-2075, 1981.

Frederick, J. H.   Commercial air transportation (5th Ed.).   Homewood, IL:
    Richard D. Irwin, Inc., 1961.

French, R. S.   The K-System MAC-1 Trouble Shooting Trainer II: Derivation
    of training characteristics.   Lowry Air Force Base, CO, AFPTRC-TN-56-9,
    1956.

Funk and Wagnalls New Standard Dictionary of the English Language.   New
    York: Funk and Wagnalls, 1963.

Gagne, R. B., Foster, H., & Crowley, M. E.   The measurement of transfer of
    training.   Psychological Bulletin, 1948, 45, 97-130.

Gaudry, E., & Spielberger, C. D.   Anxiety and educational achievement.
    Sydney: Wiley & Sons Australasia, 1971.

Gauger, K.   Analysis of the Federal Aviation Administration's Life Experi-
    ence Questionnaire.   Athens, Georgia: Institute for Behavioral Re-
    search, University of Georgia, March 1980.

General Accounting Office.   Federal employment examinations: Do they achieve
    equal opportunity and merit principle goals? Washington, D.C., FPCD-79-
    46, 1979.

Ghiselli, E. E. Dimensional problems of criteria. <u>Journal of Applied Psychology</u>, 1956, <u>40</u>, 1-4.

Gilbert, G. A. <u>Air Traffic Control: The uncrowded sky</u>. Washington, D.C.: The Smithsonian Institution Press, 1973.

Goldberg, L. R. Man versus model of man: A rationale, plus some evidence for a method of improving on clinical inferences. <u>Psychological Bulletin</u>, 1970, <u>73</u>, 422-432.

Goldberg, L. R. A historical survey of personality scales and inventories. In P. McReynolds (Ed.), <u>Advances in psychological assessment</u> (Vol. 2). Palo Alto: Science and Behavior Books, 1971.

Goldberg, L. R. Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. <u>Multivariate Behavioral Research Monographs</u>, 1972, <u>7</u>, No. 2.

Golden, C. J. Cross-cultural second order factor structure of the 16PF. <u>Journal of Personality Assessment</u>, 1978, <u>42</u>, 167-170.

Gordon, L. V. Clinical, psychometric, and work-sample approaches in the prediction of success in Peace Corps training. <u>Journal of Applied Psychology</u>, 1967, <u>51</u>, 111-119.

Great Britain Department of Industry. <u>Preliminary study of long-term air traffic systems in Europe</u> (Vol. 1). London: Her Majesty's Stationery Office, August 1977.

Greenberger, M. H., & Ward, J. H., Jr. <u>An iterative technique for multiple correlation analysis</u>. New York: International Business Machines Company, IBM Tech Newsletter No. 12, 1956.

Groenewege, A. D., & Heitmeyer, R. <u>Air freight: Key to greater profit</u>. Middlesex, England: Aerad, 1964.

Gronau, R. <u>The value of time in passenger transportation: The demand for air travel</u>. New York: Columbia University Press, 1970.

Guba, E. G. Criteria for assessing the trustworthiness of naturalistic inquiries. <u>Educational Communication and Technology</u>, 1981, <u>29</u>, 75-92.

Guilford, J. P., & Lacey, J. L. (Eds.) <u>Printed classification tests</u>. Washington, D.C.: U.S. Government Printing Office, AAF Aviation Psychology Research Program Reports, No. 5, 1947.

Guion, R. M. <u>Personnel testing</u>. New York: McGraw-Hill, 1965.

Guion, R. M. Recruiting, selection, and job replacement. In M. D.
    Dunnette (Ed.) Handbook of industrial and organizational psycho-
    logy. Chicago: Rand McNally, 1976.

Gulliksen, H., & Wilks, S. S. Regression tests for several samples.
    Psychometrika, 1950, 15, 91-114.

Hale, M. History of employement testing, In A. K. Wigdor & W. R.
    Garner (Eds.), Ability Testing: Uses, consequences and contro-
    versies. (Part II: Documentation Section). Washington, D.C.:
    National Academy Press, 1982.

Hall, C. S., & Lindzey, G. Theories of personality (2nd ed.). New
    York: John Wiley & Son, 1970.

Harrison, T. H. The International Federation of Air Traffic Con-
    trollers Associations. In A. Benoit (Ed.) A summary of modern
    air traffic control. NATO, AGARDograph, No. 209. London: Tech-
    nical Editing and Reproduction, Ltd., 1975.

Hennessy, R. T., Hockenberger, R. L., Barnebey, S. F., & Ureuls, D.
    Design requirements for an automated performance measurement
    and grading system for the UH-1 flight simulation. JSAS Catalog
    of Selected Documents in Psychology, 1981, 11, 4. (Ms No. 2177)

Henry, J. H., Kamrass, M. E., Orlansky, J., Rowan, T. C., String, J.,
    & Reichenback, R. E. Training of U.S. air traffic controllers.
    Arlington, VA: Institute of Defense Analysis, Report No. R-206,
    1975.

Holbrook, H. A. Civil aviation medicine in the bureaucracy. Chicago:
    Banner Publishing Company, 1974.

Hopkin, V. D. The psychologist's view. In D. Benoit (Ed.) A survey
    of modern air traffic control. NATA, AGARDograph No. 209. London:
    Technical Editing and Reproduction, Ltd., 1975.

Hopkin, V. D. Mental workload measurement in air traffic control. In
    N. Moray (Ed.) Mental workload, its theory and measurement. New
    York and London: Plenum Press, 1979.

Hopkin, V. D. The measurement of the air traffic controller. Human
    Factors, 1980, 22, 547-560.

Howarth, E. Were Cattell's personality sphere factors correctly identi-
    fied in the first instance? British Journal of Psychology, 1976,
    67, 213-230.

601.

Howarth, E., & Browne, J. A. An item-factor analysis of the 16PF. Personality, 1971, 2, 117-139.

Hunter, D. R., & Thompson, N. A. Pilot selection system development. JSAS Catalog of Selected Documents in Psychology, 1978, 8, 102. (Ms. No. 1790)

Hunter, J. E., & Schmidt, F. L. Fairness of selection tests, A critical analysis. Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission PS 76-5, 1976.

Institute of Social Research. National Institute for Occupational Safety and Health Report. Ann Arbor, MI: University of Michigan, 1975.

International Federation of Air Traffic Controllers Associations (IFATCA). Information Handbook. Geneva, Switzerland: 1979.

International Federation of Air Traffic Controllers Associations (IFATCA). The report of Committee C. The Controller, 3, 20, 1981.

James, L. R. Criterion models and construct validity for criteria. Psychological Bulletin, 1973, 80, 75-83.

James, L. R., Hater, L. J., Shanahan, F. M., Bruni, J. R., Jones, A., & Sells, S. B. Identification of perceived environmental factors associated with student adjustment, laboratory performance and satisfaction in the air traffic control specialist training program. Fort Worth, TX: Institute of Behavioral Research, Texas Christian University, 1980.

J. W. K. International Corporation. Study of flight service station specialists. Annandale, VA: Technical Report, October 1981.

Karson, S. Second-order factors in air traffic control specialists. Aerospace Medicine, 1967, 38, 412-414.

Karson, S. Some relations between personality factors and job performance ratings in radar controllers. Aerospace Medicine, 1969, 40, 823-826.

Karson, S., & O'Dell, J. W. Performance ratings and personality factors in radar controllers. Washington, D.C.: Federal Aviation Administration, FAA-AM-70-14, 1970.

Karson, S., & O'Dell, J. W. Is the 16PF factorially valid? Journal of Personality Assessment, 1974, 38, 104-114. (a)

Karson, S., & O'Dell, J. W. Personality differences between male and female air traffic controller applicants. Aerospace Medicine, 1974, 45, 596-598. (b)

602.

Karson, S., & O'Dell, J. W. Personality makeup of the American air traffic controller. Aerospace Medicine, 1974, 45, 1001-1007. (c)

Katzell, M. E. Expectations and dropouts in schools of nursing. Journal of Applied Psychology, 1968, 52, 154-157.

Kinney, G. C., Spahn, M. J., & Amoto, R. R. The Human element in air traffic control: Observations and analyses of the performance of the performance of controllers and supervisors in providing ATC separation services. McLean, VA: The MITRE Corporation, MTR-7655, December 1977.

Levonian, E. A statistical analysis of the 16 Personality Factor Questionnaire. Educational and Psychological Measurement, 1961, 21. 589-596.

Lewis, M. A. Objective assessment of prior air traffic control related experience through the use of the Occupational Knowledge Test. Aviation, Space, and Environmental Medicine, 1978, 49, 1155-1159.(a)

Lewis, M. A. Use of the Occupational Knowledge Test to assign extra credit in selection of air traffic controllers. Washington, D.C.: Federal Aviation Administration, Report No. FAA-AM-78-7, 1978.(b)

Lewis, M. A. A comparison of three models for determining test fairness. Washington, D.C.: Federal Aviation Administration FAA-AM-79-3, 1979.

Linn, R. L. Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), Ability testing: Uses, consequences, and controversies. (Part II: Documentation Section). Washington, D.C.: National Academy Press, 1982.

Loevinger, J. Some limitations of objective personality tests. In J. N. Butcher (Ed.), Objective personality assessment. New York: Academic Press, 1972.

Loevinger, J., Gleser, G. C., & DuBois, P. H. Maximizing the discriminating power of a multiple-score tess. Psychometrika, 1953, 18, 309-317.

Lofquist, L. H., & Dawis, R. V. Adjustment to work: A psychological view of man's problems in a work-oriented society. New York: Appleton-Century-Crofts, 1969.

Maekawa, M. Now in Japan. The Journal of Air Traffic Control, 1980, 22(2), 24-25.

Magnusson, D. Test theory. Reading, MA: Addison-Wesley, 1966.

Manchester, L. Canada's aviation industry. Toronto: McGraw-Hill, 1968.

Mathews, J. J., & Cobb, B. B. Relationships between age, ATC experience and job ratings of terminal area traffic controllers. <u>Aerospace Medicine</u>, 1974, <u>45</u>, 56-60.

Mathews, J. J., Cobb, B. B., & Collins, W. E. <u>Attitudes on en route air traffic control training and work: A comparison of recruits initially trained at the FAA Academy and recruits initially trained at assigned centers</u>. Washington, D.C.: Federal Aviation Administration, FAA-AM-75-3, 1975.

Mathews, J. J., Collins, W. E., & Cobb, B. B. <u>A sex comparison of reasons for attrition of non-journeyman FAA air traffic controllers</u>. Washington, D.C.: Federal Aviation Administration, FAA-AM-74-2, 1974. See also: A sex comparison of reasons for attrition in a male-dominated occupation. <u>Personnel Psychology</u>, 1974, <u>27</u>, 535-541. (a)

Mathews, J. J., Collins, W. E., & Cobb, B. B. <u>Job-related attitudes of non-journeyman FAA air traffic controllers and former controllers: A sex comparison</u>. Washington, D.C.: Federal Aviation Administration, FAA-AM-74-7, 1974. (b)

McGuigan, F. <u>Air traffic controller acquisition project: Review of the existing system, comparative analysis, and recommendations</u>. Civil Aeronautics, Air Traffic Service. Technical Report. Ottawa: Transport Canada Air, 1979.

McGuire, C. H., Solomon, L. M., & Bashook, P. G. <u>Construction and use of written simulations</u>. New York: The Psychological Corporation, 1976.

Meehl, P. E. <u>Clinical vs. statistical prediction</u>. Minneapolis: University of Minnesota Press, 1954.

Mies, J. M., & Colmen, J. G. Development of recommendations for ATCS selection tests. Report of Task I. Washington, D.C.: Education and Public Affairs, Inc., 1976.

Mies, J., Colmen, J. G., & Domenech, O. <u>Predicting success of applicants for positions as air traffic control specialists in the Air Traffic Service</u>. Washington, D.C.: Education and Public Affairs Inc., Contract DOT FA-75WA-3646, 1977.

Milne, A. M., & Colmen, J. G. <u>Selection of air traffic controllers for the Federal Aviation Administration</u>. Washington, D.C.: Federal Aviation Administration, Office of Aviation Medicine. Final Report, Contract No. DOT-FA70WA-2371, 1972.

Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M., Review and conceptual analysis of the employee turnover process. <u>Psychological Bulletin</u>, 1979, <u>86</u>, 493-522.

Mohler, S. R. The air traffic controller: Health and safety. Human Factors Bulletin, Flight Safety Foundation, Arlington, VA, April 1980.

Muchinsky, P. M., & Tuttle, M. L. Employee turnover: An empirical and methodological assessment. Journal of Vocational Behavior, 1979, 14, 43-77.

Nagay, J. A. Field tryout of a procedure for evaluating the proficiency of air route traffic controllers. Washington, D.C.: Civil Aeronautics Administration, Division of Research, Report No. 91, October 1950.

National Transportation Safety Board. Air traffic control system. Washington, D.C.: Department of Transportation, Special Investigation Report, NTSB-SIR-81-7, December 1981.

Nealey, S. M., Thornton, G. C., III, Maynard, W. S., & Lindell, M. K. Defining research needs to insure continued job motivation of air traffic controllers in future air traffic control systems. Seattle: Battelle Memorial Institute, Human Affairs Research Centers, Final Report of Contract No. DOT-FAA74WA1-499, 1975.

Novick, M. R. Ability testing: Federal guidelines and professional standards. In A. K. Wigdor & W. R. Garner (Eds.), Ability testing: Uses, consequences, and controversies. (Part II: Documentation Section). Washington, D.C.: National Academy Press, 1982.

Older, H. J., & Cameron, B. J. Human factors aspects of air traffic control. Washington, D.C.: NASA Report No. CR-1957, 1972.

Olofsson, S. The use of simulation at the Swedish Air Traffic Services Academy (SATSA). The Journal of Air Traffic Control, 1980, 22, 26-29.

Owens, W. A. Background data. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Peterson, P. B., & Lippitt, G. L. Comparison of behavioral styles between entering and graduating students in Officer Candidate School. Journal of Applied Psychology, 1968, 52, 66-70.

Pickrel, E. W. Development of occupational knowledge tests for en route, terminal, and flight service station air traffic control specialists. Unpublished document, FAA Office of Aviation Medicine, 1977.

Pickrel, E. W. Performance standards for pass-fail determinations in the national air traffic flight service station training program. Washington, D.C.: Federal Aviation Administration, FAA-AM-79-18, 1979.

605.

Potkay, C. R. The role of personal history data in clinical judgment: A selective focus. Journal of Personality Assessment, 1973, 37, 203-213.

Rock, D. B., Dailey, J. T. Ozur, H., Boone, J. O., & Pickrel, E. W.. Selection of applicants for the air traffic controller occupation. Washington, D.C.: Federal Aviation Administration, Report No. FAA-AM-82-11, 1982.

Ronan, W. W., & Prien, E. P. Towards a criterion theory: A review and analysis of research and opinion. Greensboro, N.C.: The Richardson Foundation, 1966.

Rose, R. M., Jenkins, C. D., & Hurst, M. W. Air traffic controller health change study. Washington, D.C.: Federal Aviation Administration, FAA-AM-78-39, 1978.

Saleh, S. D., Lee, R. J., & Brien, E. P. Why nurses leave their jobs-- an analysis of female turnover. Personnel Administration, 1965, 28, 25-28.

Schmidt, F. L., & Hunter, J. E. Development of a general solutions to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. Validity generalization results for two job groups in the petroleum industry. Journal of Applied Psychology, 1981, 66, 261-273.

Sealy, K. R. The geography of air transport. London: Hutchinson University Library, 1957.

Sells, S. B. On the nature of stress. Chapter 9 in J. E. McGrath (Ed.), Social and psychological factors in stress. New York: Holt, Rinehart, and Winston, Inc., 1970.

Sells, S. B., Demaree, R. G., & Will, D. P. Dimensions of personality: I. Conjoint factor structure of Guilford and Cattell trait markers. Multivariate Behavioral Research, 1970, 5, 391-422.

Sells, S. B., Demaree, R. G., & Will, D. P. Dimensions of personality: II. Separate factor structure in Guilford and Cattell trait markers. Multivariate Behavioral Research, 1971, 6, 135-186.

Sells, S. B., Demaree, R. G., & Will, D. P. A taxonomic investigation of personality. Conjoint factor structure of Guilford and Cattell trait markers. Fort Worth, TX: Institute of Behavioral Research, Texas Christian University, Final Report: U.S. Office of Education, Contract No. OE-t-10-296, 1968.

Siegel, P. V. The psychological makeup of the air traffic controller. Paper presented at the International Meeting on Aerospace Medicine, Sydney, Australia, November 29, 1966.

Smith, P. C. Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.) Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Smith, R. C. Comparison of the job attitudes and interest patterns of air traffic and airway facility personnel. Aviation, Space, and Environmental Medicine, 1979, 50, 1031-1036.

Smith, R. C. Stress, anxiety and the air traffic control specialist: Some conclusions from a decade of research. Washington, D.C.: FAA Office of Aviation Medicine, FAA-AM-80-14, 1980.

Smith, R. C., & Hutto, G. L. Vocational interests of air traffic control personnel. Aviation, Space, and Environmental Medicine, 1975, 46, 871-877.

Solberg, C. Conquest of the skies: A history of commercial aviation in America. Boston: Little, Brown, & Co., 1979.

Spahn, M. J. The human element in air traffic control: Observations and analyses of the performance of controllers and supervisors in providing ATC separation services. Supplement 2: Analysis of the system effectiveness information system data base. McLean, VA: The MITRE Corporation, MTR-7655, Supplement 2, December 1977.

Spielberger, C. D., Gorsuch, R, & Lushene, R. State-trait anxiety inventory, preliminary test manual. Palo Alto: Consulting Psychologists Press, 1969.

Stammers, R. B. Human factors in airfield air traffic control. Ergonomics, 1978, 21, 483-488.

Sullivan, E. T., Clark, W. W., & Tiegs, E. W. Manual for California Test of Mental Maturity, Advanced Form. Los Angeles: California Test Bureau, 1957.

Sundberg, N. D. Assessment of persons. Englewood Cliffs, NJ: Prentice-Hall, 1977.

System Development Corporation. Air traffic controller job task analysis. Santa Monica, CA, Report No. TM(L) - 4925/000/00, 1972.

System Development Corporation. Phase II - Local controller, ground controller, radar controller report. Santa Monica, CA, Report TM(L) - 4925/000/00, 1972.

System Development Corporation. FAA tower (CAB) descriptions and flow diagrams of control functions. Santa Monica, CA, Report TM-5271/001/00, 1974.

System Development Corporation. Air route traffic control center - Descriptions and flow diagrams of control functions. Santa Monica, CA, Report TM-5329/000/00, 1974.

System Development Corporation. Air route traffic control center - Controller over-the-shoulder training review instruction manual. Santa Monica, CA, Report TM-5330/000/00, 1974.

System Development Corporation. Air route traffic control center - Controller extended performance rating instruction manual. Santa Monica, CA, Report TM-5331/000/00, 1974.

System Development Corporation. Terminal radar approach control facility (Tracon) descriptions and flow diagrams of control functions. Santa Monica, CA, Report TM-5466/000/00, 1975.

System Development Corporation. Terminal Area (Tracon/CAB) Air traffic control facility controller training review instruction manual. Santa Monica, CA, Report TM-5485/000/00, 1975.

System Development Corporation. Terminal option controller performance evaluation report. Santa Monica, CA, Report TM-5491/000/00, 1975.

System Development Corporation. Terminal area (Tracon/CAB) air traffic control facility - Controller performance rating instruction manual. Santa Monica, CA, Report TM-5493/000/00, 1975.

Taylor, J. A. A personality scale of manifest anxiety. Journal of Abnormal Social Psychology, 1953, 48, 285-290.

Taylor, M. V., Jr. The development and validation of a series of aptitude tests for the selection of personnel for positions in the field of air traffic control. Pittsburgh, PA: American Institute for Research, August 1952.

Tenopyr, M. L., & Oeltjen, P. D. Personnel selection and classification. Annual Review of Psychology, 1982, 33, 581-618.

608.

Thackray, R. J. Boredom and monotony as a consequence of automation: A consideration of the evidence relating boredom and monotony to stress. Washington, D.C.: FAA Office of Aviation Medicine, FAA-AM-80-1, February 1980.

Thackray, R. J. The effect of age on the ability to sustain attention during performance of a simulated radar task. Oklahoma City, OK: FAA Civil Aeromedical Institute, Research Task, Bimonthly Report, Task No.: AM-C-81PSY-84, Aviation Psychology Laboratory, April 1981.

Thackray, R. J., & Touchstone, R. M. The effect of age on complex monitoring performance. Oklahoma City, OK: Laboratory Note, Aviation Psychology Laboratory, FAA Civil Aeromedical Institute, 1980. (a)

Thackray, R. J., & Touchstone, R. M. An exploratory investigation of various assessment instruments as correlates of complex visual monitoring performance. Washington, D.C.: FAA Office of Aviation Medicine, FAA-AM-80-17, October 1980. (b)

Thorndike, R. L. Personnel selection: Test and measurement technique. New York: Wiley, 1949.

Toops, H. A. The use of addends in experimental control, social census and managerial research. Psychological Bulletin, 1948, 45, 41-74.

Trites, D. K. Problems in air traffic management: I. Longitudinal prediction of effectiveness in air traffic controllers. Aerospace Medicine, 1961, 32, 1112-1118. See also: Problems in air traffic management: I. Longitudinal prediction of effectiveness of air traffic controllers. Washington, D.C.: Federal Aviation Administration, FAA-AM-61-1, 1961.

Trites, D. K. Problems in air traffic management: VI. Interaction of training-entry age with intellectual and personality characteristics of air traffic control specialists. Aerospace Medicine, 1964, 35, 1184-1194.

Trites, D. K. Problems in air traffic management: VI. Interaction of training entry age with intellectual and personality characteristics of air traffic control specialists. Washington, D.C.: Federal Aviation Administration, FAA-AM-65-21, 1965.

Trites, D. K., & Cobb, B. B. Problems in air traffic management: IV. Comparison of pre-employment job-related experience with aptitude test predictors of training and job performance of air traffic control specialists. Washington, D.C.: Federal Aviation Administration, FAA-AM-63-31, 1963.

Trites, D. K. & Cobb, B. B. CARI research on air traffic control
    specialists: Age, aptitude, and experience as predictors of
    performance. Oklahoma City, OK: FAA Civil Aeromedical Research
    Institute, Unnumbered Report, 1964.

Trites, D. K., & Cobb, B. B. Problems in air traffic management: III.
    Implications of training-entry age for training and job performance
    of air traffic control specialists. Aerospace Medicine, 1964, 35,
    336-340. (b)

Trites, D. K., & Cobb, B. B. Problems in air traffic management: IV.
    Comparison of pre-employment, job-related experience with aptitude
    tests as predictors of training and job performance of air traffic
    control specialists. Aerospace Medicine, 1964, 35, 428-436. (a)

Trites, D. K., Kurek, A., & Cobb, B. B. Personality and achievement
    of air traffic controllers. Aerospace Medicine, 1967, 38, 1145-1150.

Trites, D. K., Miller, M. C., & Cobb, B. B. Problems in air traffic
    management: VII. Job and training performance of air traffic
    control specialists - Measurement, structure, and prediction.
    Aerospace Medicine, 1965, 36, 1131-1138. See also: Problems in
    air traffic management: VII. Job and training performance of air
    traffic control specialists - Measurement, structure, and predic-
    tion. Washington, D.C.: Federal Aviation Administration, AM-65-
    22, 1965.

Tucker, J. A. Relative predictive efficiency of multiple regression
    and unique pattern techniques. (Doctoral dissertation, Columbia
    University, 1950). Microfilm Abstracts, 1951, 11(s), 489-440.

Tucker, J. A. Simulating the simulator. Proceedings of the 1981 Summer
    Computer Simulation Conference, 1981, 501-502.

Uniform guidelines on employee selection procedures. Federal Register,
    1978, 43, (166):38295-38310.

Van Deventer, A. D. Biographical profiles of successful and unsuccessful
    air traffic control specialist trainees. Paper presented at the
    Annual Meeting of the Aerospace Medical Association, Anaheim, CA,
    May, 1980.

Vaughn, D. S. The relative methodological soundness of several major
    personality factor analyses. Journal of Behavioral Science, 1973,
    1, 305-313.

Wallace, S. R. Criteria for what? American Psychologist, 1965, 20, 411-417.

Washington Post. Air traffic seems to run smoothly, safely 12 months later.
    August 3, 1982.

Whitfield, D., & Stammers, R. B. The air traffic controller. In W. T. Singleton (Ed.) The study of real skills Volume I, The analysis of practical skills. Lancaster, England: MTP press, 1978.

Whitnah, D. R. Safer skyways: Federal control of aviation (1926-1966). Ames, IA: Iowa State University Press, 1966.

Wigdor, A. K. & Garner, W. R. (Eds.) Ability testing: Uses, consequences, and controversies. (Part I: Report of the Committee). Washington, D.C.: National Academy Press, 1982.

Worker attributes for air traffic control jobs and their measurement. Washington, D.C.: Education and Public Affairs, Inc., 1970.

END

FILMED

1-85

DTIC